Princess Sumaya University for Technology
Since 1991 تأسست
PSUT
جامعة الأميرة سمية للتكنولوجيا

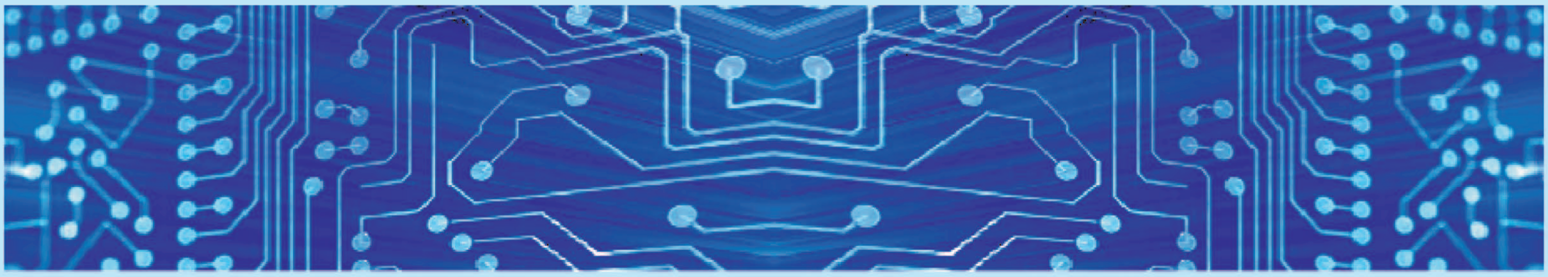
صندوق دعم البحث العلمي
Scientific Research Support Fund

Jordanian Journal of Computers and Information Technology

December 2019 VOLUME 05 NUMBER 03
ISSN 2415 - 1076 (Online)
ISSN 2413 - 9351 (Print)

JJCIT

www.jjcit.org jjcit@psut.edu.jo



An International Peer-Reviewed Scientific Journal
Financed by the Scientific Research Support Fund

Jordanian Journal of Computers and Information Technology (JJCIT)

The Jordanian Journal of Computers and Information Technology (JJCIT) is an international journal that publishes original, high-quality and cutting edge research papers on all aspects and technologies in ICT fields.
JJCIT is hosted by Princess Sumaya University for Technology (PSUT) and supported by the Scientific Research Support Fund in Jordan. Researchers have the right to read, print, distribute, search, download, copy or link to the full text of articles. JJCIT permits reproduction as long as the source is acknowledged.

Aims and Scope

The JJCIT aims to publish the most current developments in the form of original articles as well as review articles in all areas of Telecommunications, Computer Engineering and Information Technology and make them available to researchers worldwide. The JJCIT focuses on topics including, but not limited to: Computer Engineering & Communication Networks, Computer Science & Information Systems and Information Technology and Applications.

Indexing

JJCIT is indexed in:

ELSEVIER Scopus

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

crossref Text and Data Mining

OCLC WorldCat

BASE

ACADEMIA

Google scholar

JJCIT Address

Website: www.jjcit.org
Email: jjcit@psut.edu.jo
Address: Princess Sumaya University for Technology, Khalil Saket Street, Al-Jubaiha
B.O. Box: 1438 Amman 11941 Jordan
Telephone: +962-6-5359949
Fax: +962-6-7295534

EDITORIAL BOARD

INTERNATIONAL ADVISORY BOARD

“Opinions or views expressed in papers published in this journal are those of the author(s) and do not necessarily reflect those of the Editorial Board, the host university or the policy of the Scientific Research Support Fund”.

"ما ورد في هذه المجلة يعبر عن آراء الباحثين ولا يعكس بالضرورة آراء هيئة التحرير أو الجامعة أو سياسة صندوق دعم البحث العلمي".

EVALUATING THE ACCESSIBILITY OF KUWAITI E-GOVERNMENT WEBSITES

Iyad Abu Doush[1] and Zainab AlMeraj[2]

(Received: 8-May-2019, Revised: 21-Jun.-2019 and 18-Jul.-2019, Accepted: 28-Jul.-2019)

ABSTRACT

Nowadays, the Web is used as a medium for providing different kinds of services for people. The needs of people with disabilities have to be taken into consideration when developing E-government. The objective of this work is to evaluate the Web accessibility issues facing people with disabilities in Kuwait in an attempt to identify problems, enhance government awareness and promote inclusion. In order to evaluate the accessibility of E-government services in Kuwait, we apply automatic and expert evaluation on the top 17 E-government services used in 2018 in Kuwait. The obtained results show that 13 of the evaluated E-services are impossible to use and thus reveal a serious weakness in adhering to the Web Content Accessibility Guidelines (WCAG) 2.0 level A for most of the evaluated websites. In addition, the study shows the importance of following the task-based approach when evaluating the accessibility of the websites, as navigation between different pages when completing each task can help in discovering other accessibility issues.

KEYWORDS

E-government Services, Accessibility, Blind, People with Disabilities, Web Accessibility Evaluation.

1. INTRODUCTION

Web accessibility means that people with disabilities can perceive, understand, navigate, interact with and contribute to the Web regardless of age or ability [1]. Accessibility encompasses all disabilities that affect access to the Web, including visual, auditory, physical, speech, cognitive and neurological disabilities.

The increased use of Information and Communication Technologies (ICT) and the affordable Internet access has engaged governments around the world to provide their citizens, residents, visitors and businesses with online services through government portal [2]. According to Taewoo Nam [3], there are five types of E-government uses which are: service use, general information use, policy research, participation and co-creation. The world development report [4] mentions that the benefits of providing E-government services include stable relations with citizens, better delivery, savings and more efficiency.

The World Wide Web Consortium (W3C) is an organization for standardization of the Web. W3C developed a set of accessibility recommendations called Web Content Accessibility Guidelines (WCAG 2.0). These guidelines provide a set of recommendations for Web developers to provide a Web content which can be accessed by people with different kinds of disabilities. Another set of recommendations and guidelines is introduced in Section 508. It demands that the electronic information for the US government must be accessible to people with disabilities.

To achieve universal digital access, the Web Accessibility Initiative (WAI) established a set of guidelines to help make the Web accessible for people with disabilities, later called the Web Content Accessibility Guidelines (WCAG) in 1999. Since then, WCAG have been used as standards for developing accessible Web content. The standards start with version 1.0 (i.e., WCAG 1.0) which focuses on HTML. After that, the standard was updated in 2008 to WCAG 2.0 by focusing on generic digital assets instead of a specific technology. Lately in 2018, WCAG 2.1 was introduced to add success criteria that are not in WCAG 2.0. As mentioned by W3C[1], "WCAG 2.1 does not deprecate or supersede WCAG 2.0".

[1] https://www.w3.org/WAI/standards-guidelines/wcag/

1. I. Abu Doush is with Department of Computer Science and Information Systems, American University of Kuwait, Salmiya, Kuwait and Department of Computer Sciences, Yarmouk University, Irbid, Jordan. E-mail: idoush@auk.edu.kw
2. Z. AlMeraj is with Department of Information Science, Kuwait University, Adailiya, Kuwait. E-mail: z.almeraj@gmail.com

The countries across the globe provide E-government to enhance transparency and allow faster and easier interactions between citizens. However, the efficiency and accessibility of these E-government services have been investigated to provide equal access to all country citizens [2]. Based on the United Nations E-government report, Kuwait is ranked 40th out of 193 countries in the development index and 55th out of 193 countries in the E-participation index.

The State of Kuwait launched the Kuwait Government Online portal (https://www.e.gov.kw) through the Central Agency for Information Technology (CAIT) to provide information and services to all citizens, residents and visitors in addition to the governmental and business sectors. The portal is reachable through the Web or mobile platforms and lists usage statistics, news, announcements, laws and regulations. Since its establishment, the government has been committed to achieve a usable working government portal to encourage citizens' participation and to empower all those living and residing in Kuwait. The government has been further motivated to achieve the realization of efficient E-government to fulfil the New Kuwait Vision 2035, announced in 2016, which entails areas contingent on ICT and cloud-based technologies [5].

Previous studies have shown that accessibility in E-government in the Middle East and the Gulf Cooperation Council (GCC) region suffers from several weaknesses. The main challenge facing the E-governments in Kuwait, Qatar and UAE is the lack of knowledge of Web accessibility standards among Web developers [6]. Saleem [6] tested 10 different sites and checked their conformance to the accessibility standard (WCAG 2.0). The results show after testing five pages from 10 Kuwaiti government websites that the percentage of error is 93%.

The total number of registered people with disabilities in Kuwait in 2016 is 51,243[2]. The inclusion of people with disabilities by offering them different services is essential to ensure their financial security and engagement with society, as well as to promote independence [7]. Kuwait is among the countries which signed the UN convention on the Rights of Persons with Disabilities (CRPD) [8]. In addition, the country has its own disability law to protect and support people with disabilities [9].

Kuwait is currently working on achieving Kuwait 2035 Vision through the ratification of the UN CRPD for people with disabilities and achieving 17 of the UN Sustainable Development Goals (SDG) [10] through the following:

1. Enhancing human capacities and institutional effectiveness for prevention, early detection, diagnosis and rehabilitation of disabilities.
2. Removing barriers to social, economic and educational inclusion of people with disabilities.
3. Increasing technical expertise and organizational capacities for the implementation of universal design and countrywide use of technology enablers [11].

In pursuit to achieve this vision, the General Secretariat of the Supreme Council for Planning and Development (GSSCPD) in collaboration with Public Authority for Disabled Affairs (PADA) and the UNDP have publicly announced the Kuwait National Framework for Digital Accessibility in May 2018 based on international standards with the aim of transforming the Kuwaiti digital environment into a qualified and supportive environment for people with disabilities [10].

This paper evaluates a set of commonly used government E-services using two approaches; multiple automatic tools and task-based expert reviews. The results will further help in promoting equality, inclusion, awareness, learning and implementation of local and international accessibility standards.

In this study, a task-based approach is used in which different pages are navigated in order to evaluate the accessibility of the provided E-service. This help in checking the accessibility of the provided E-service in a situation similar to what the user encounters in real life. Furthermore, this enables checking the ease of navigation between the pages when the user is completing the service.

The rest of the paper is organized as follows. We first give a brief overview of related work (Section 2). We then present our methodology (Section 3). After that, we present and discuss the findings (Section 4). Finally, we present conclusions and discuss opportunities for future work (Section 5).

[2] http://www.pada.gov.kw/

2. RELATED WORK

Government websites enable citizens to easily interact with them through efficient platforms that host public information and services. To access these E-services, people with disabilities use computers with specialized software commonly called assistive technologies. Screen readers are the most popular type of assistive technology for users with visual impairments [12]. Those with hearing impairments, cognitive disabilities and motor skill impairments may require other technologies, such as voice browsers, special joysticks or trackballs [49].

Many studies have investigated the accessibility of government websites. One form of evaluating accessibility involves benchmarking website designs and functionality with WCAG 2.0 guidelines. These guidelines and success criteria are organized around four principles to allow access and use of Web content [13]:

1. Perceivable - Information and user interface components must be presentable to users in a way that they can perceive. This means that users must be able to acquire the presented information through their senses.
2. Operable - User interface components and navigation must be operable. This means that users must be able to operate the interface (i.e., the interface cannot require interaction that a user cannot perform).
3. Understandable - Information and the operation of user interface must be understandable. This means that users must be able to understand the information as well as the operation of the user interface. For example, the information sequencing is meaningful to the user or the presented information allows the user to complete the required action.
4. Robust - Content must be robust enough so that it can be interpreted reliably by a wide variety of user agents, including assistive technologies.

The most common means of accessibility evaluation is the use of automatic tools to check whether or not the Web pages follow the WCAG 2.0 guidelines. The most popular automatic tools include AChecker, HTML Validator, CSS Validator, APrompt, Cynthia, Says and EvalAccess 2.0 [50]. These tools differ in their criteria from effciency to conformance levels (A, AA and AAA). In WCAG, the success criteria for level A are easy to meet and do not affect the website design or structure. On the other hand, levels AA and AAA are more strict and require more work [51]. Because of this, we selected level A in our evaluation which represents the minimum level of accessibility requirements.

In an automatic evaluation study [14], the authors analyzed the usability, accessibility and vulnerability of 61 Turkish E-government websites using six automatic tools that assess navigation, HTML errors, content quality, conformance to W3C standard and compatibility. Their results show that E-government websites did not conform to international standards despite efforts made towards implementing a Web accessibility policy issued by the government in 2001. It is important to improve design features of E-government websites to be more effective and user-centric [14].

Similar studies have been conducted with variations of automatic tools for multiple governments, such as Ghana by Yaokumah et al. [15], Kenya by Wanyonyi Kituyi and Waweru [16], Saudi Arabia by Al-Faries et al. [17], Jordan by Doush et al., among others [18]-[19], [52]-[53], Dubai by Mourad and Kamoun [20] and Kamoun and Almourad [21] and Korea by Park [22] with similar outcomes suggesting the need to improve website accessibility. A cross-continent study by Patr et al. [23] evaluated 15 Asian government websites and found overwhelming evidence of lack of accessibility awareness and implementation.

In a multilingual study [24], ten E-government websites from the Arab world were evaluated in English and Arabic using four automatic evaluation tools; SortSite, TAW, AChecker and WAVE. Its findings include noticeable un-explained differences between accessibility scores of Arabic *versus* English websites. The authors discussed the importance of choosing the right tool for the right evaluation test and the adjustment of government regulations to include rules and guidelines for developers and managers.

A total of 302 Indian universities' websites are evaluated using automatic accessibility evaluation tools in [25]. After that, the websites are classified into poorly accessible websites, intermediately accessible websites and highly accessible websites.

In an effort to check enhances in conformance levels with WCAG 2.0, Al-Khalifa [26] evaluated the accessibility of the Arabic version of 34 Saudi government websites in 2010 and re-evaluated them in 2016 [27]. Each homepage was inspected using the WAVE and Web developer evaluation toolbars. In the 2010 evaluation, it was found that no website followed the minimum guidelines set by WCAG 2.0. But after enforced regulations on E-readiness and accessibility in 2016, remarkable improvements were noticed.

Some accessibility evaluators tested accessibility using automatic quantitative metric scores such as the work conducted by Vigo et al. [28]. They tested seven website accessibility metrics and showed that three of them; namely, accessibility quantitative metric, page measure and Web accessibility barrier produce higher quality assessments than the others. Relying solely on these kinds of scores does pose an issue for evaluation as we discuss later.

Akram and Sulaiman [29] discussed accessibility issues on Saudi Arabia E-government websites. The authors identified Web accessibility issues of government and university sites across the country and stated that it is not enough to just check WCAG 2.0 conformance and that more effective standards need to be identified. According to AkgUL and Vatansever [30], human judgment is needed to provide an accurate evaluation of Web accessibility. Automatic tool evaluations, for example, cannot give a full picture of the interaction between Web contents and assistive technology; they cannot detect all violations and can therefore result in false positives and false negatives [21].

Other forms of accessibility evaluations include expert reviews, end-user testing, Web developer surveys and combinations of the latter commonly referred to as multi-method approaches. In a multi-method approach [31], the authors provided an accessibility evaluation of 100 federal home pages using both human and automated methods to check conformance to Section 508 [32] accessibility guidelines. Using two methods, they found a better understanding of the limitations and suggested improving policy related to Section 508 compliance for websites for better accessibility.

Another study by Jaeger [33] used a different multi-approach method which combines policy analysis, expert testing, user testing, automated testing and Web developer questionnaires with the aim of using each result to present an accurate presentation of the accessibility status.

In the Malaysian study by Hanapi [34], one automatic tool and a survey were adopted to understand E-government Web developers' awareness of accessibility guidelines. More recently, Doush et al. [18] performed three evaluation methods to get a deeper understanding of how to enhance accessibility in Jordan. They recruited 20 blind users to test the accessibility of nineteen E-government websites, distributed a survey to E-government Web developers and recruited two experts to test all websites for a complete assessment. Their findings suggest a lack of awareness, understanding and adoption of accessibility guidelines.

E-government websites in three Gulf Cooperation Council (GCC) countries: UAE, Kuwait and Qatar have been studied by Saleem [6]. The author used case studies, automated website assessments, manual assessments and document analysis. He found that the Arabic language accessibility resources and tools were causing limitations in providing accessibility. In a follow-up study by Saleem [35], he went on to further develop and implement an Arabic accessibility resource of guidelines for Arabic speaking Web developers. A similar approach was used for localizing Web accessibility content of Arabic university websites in Saudi Arabia by [36].

E-government evaluations have also brought together testing of accessibility and usability of websites. The investigation by Al-Faries et al. [17] evaluated the accessibility of 20 E-government websites using automatic tools and went on to evaluate the usability of the same websites using expert reviews. In this work, we conduct a multi-method approach using five automatic tools followed by expert reviews in an effort to identify the extent of accessibility problems facing people with disabilities when searching for information and E-services on the Kuwait government portal.

3. METHODOLOGY

3.1 E-government Web Site Selection

There are sixty-one organizations listed on the Kuwait E-government website portal as of May 2018 [37]. The total number of E-government services offered was 1902 services which are splitted between

909 E-services and 993 informative services. The websites selected for the study were chosen from a list of the top 20 E-services in 2018 collected from the Central Agency for Information Technology[3]. In Kuwait, each government agency has its own website with same E-service that Kuwaiti government online portal provides, but in different style and format. The E-services in the government agency websites have not been tested.

The top 17 E-services in the list are used as a sample for the Kuwaiti E-government evaluation (see Table 1). Note that N/A in the table is for the unavailable tasks which are 3, 6, 10, 11, 15, 16 and 17. This is because either the service was unavailable (3, 6, 11 and 16) or the service link was not found (10, 15 and 17) using the search box or by looking into the service links in the government portal homepage.

There are mainly five types of E-government services; service use, general information use, policy research, participation in decision-making and co-creation of policies, information and services with government and other citizens [3]. The type of the service is listed in the table as well as the number of tested pages. The number of tested pages represents how many pages are needed to complete the service by the expert when evaluating the service.

In addition to these E-services, we evaluated the homepage of the Kuwaiti government online portal[4]. The English version of each of these websites was chosen for the evaluation as a preliminary step in this investigation. Future work will include evaluating the Arabic versions.

Table 1. The top seventeen E-services of the Kuwaiti E-government for the year 2018, highlighted rows are for unavailable services.

No.	E-service	Government agency	No. of tested pages
1	Inquiring about Civil ID Status	Public Authority for Civil Information	2
2	Violation Payment (Traffic & Immigration)	Ministry of Interior	3
3	Inquiring about lawsuits filed against you	Ministry of Justice	N/A
4	Inquiring about phone bill and E-payment	Ministry of Communication	4
5	Electricity and water bills enquiry and E-payment	Ministry of Electricity and Water	4
6	Renew work permit	Public Authority of Manpower	N/A
7	Request appointment for food checkup	Ministry of Health	3
8	Mobile bill payment and recharge services	eNet Company	4
9	Results of staff inward and outward transfer	Ministry of Education	3
10	Inquiry into status of an application (altarasul system)	Civil Service Commission	N/A
11	Reserve a hall	Ministry of Social Affairs and Labor	N/A
12	Multi-civil renewal and payments	Public Authority for Civil Information	3
13	Inquiring about travel ban	Ministry of Interior	1
14	Personal inquiry about MOI	Ministry of Interior	4
15	Civil ID fines E-payment	Public Authority for Civil Information	3
16	Inquiring about travel violations	Ministry of Interior	3
17	Inquiring about arrest warrants	Ministry of Justice	N/A

[3] We would like to thank Mrs. Nadia AlKhalifa, Statistics Department, CIAT for providing this information.
[4] https://www.e.gov.kw/sites/kgoEnglish/Pages/HomePage.aspx

3.2 E-government Website Accessibility Evaluation

In order to evaluate the Web accessibility of the selected Kuwaiti E-government services, we used multiple automatic tools and expert reviews. In addition, we tested the conformance of the E-services to the HTML and CSS standards. Note that the Kuwaiti E-government services are evaluated using a personal computer (PC).

3.2.1 Automatic Tools

As mentioned in Section 2, there are several automatic tools that can be used for Web accessibility evaluation. We selected three automatic evaluation tools; AChecker, Total Validator and WAVE jointly to overcome any drawbacks of a single tool use, as mentioned in [38]. The study by Bazeem et al. [39] investigated 23 evaluation methods and favored the results of the Web Accessibility Checker (AChecker) over the other tools.

We use AChecker [40] to automatically evaluate the selected E-government websites. AChecker can be used to check the website conformance to standards and guidelines, such as WCAG 1.0, WCAG 2.0 and Section 508. The tool classifies the recognized problems into the following: known problems (these are certain accessibility barriers), likely problems (these are probably accessibility barriers), and potential problems (these need a human decision). In order to share accurate results, we only presented the known problems detected for WCAG 2.0 with a level of conformance and left the others for expert interpretation.

The second tool used in our evaluation is Total Validator [41] with the basic feature settings. The tool can check the accessibility against the standards WCAG 1.0, WCAG 2.0 and Section 508 as shown previously in [27]. Lastly, we determine the level of Accessible Rich Internet Applications (ARIA) standards usage of the evaluated websites using a third online tool called WAVE [42]. WAVE is utilized to find the number of ARIA features in the evaluated website, errors, contract errors and structural elements. Note that in the automatic tool evaluations the first Web page of the used service is evaluated and it is the E-service URL shown in Table 3.

3.2.2 HTML and CSS Validation

HTML and CSS code validation refers to comparing Web page scripts against syntax rules and standard specifications. The validation of hypertext markup language (HTML) is considered one of the main steps of evaluating Web accessibility according to researchers [43]. Assistive technologies rely on these standards when accessing HTML and cascading style sheets (CSS) [44].

In addition to the automatic evaluation tools, we validate the websites selected using the HTML markup validation service[5] and CSS validator service[6]. These two services are available for free from World Wide Web Consortium (W3C).

3.2.3 Expert Reviews

Many of the accessibility problems cannot be identified using automatic tools (e.g., inaccessible Captcha). In order to investigate the accessibility problems deeply, we use expert evaluation. Expert evaluation can be applied by examining the code of the Web page to look for accessibility problems or the evaluator can simulate the user usage of the E-service by completing a specific task. In this study, we apply a task-based expert review to mimic the challenges encountered by the users when they use the E-services. The experts are the two authors of the paper who have expertise in Web accessibility guidelines.

The evaluated tasks are shown in Table 1. The experts did a walk-through of each website to complete the defined tasks in a similar process adopted by the experts' review in [18]. In order to identify the accessibility problems in the website, the tasks are completed using NVDA screen reader[7] using the English language. NVDA screen reader is selected to perform the evaluation, because it is free and it is one of the most popular screen readers [48].

[5] https://validator.w3.org/

[6] https://jigsaw.w3.org/css-validator/

[7] https://www.nvaccess.org/

Table 2. The success criteria for level A used in the manual evaluation.

No.	Success criteria	Criteria (all level A)	No.	Success criteria	Criteria (all level A)
1	1.1.1	Non-text Content	14	2.3.1	Three Flashes or Below Threshold
2	1.2.1	Audio-only and Video-only (Prerecorded)	15	2.4.1	Bypass Blocks
3	1.2.2	Captions (Prerecorded)	16	2.4.2	Page Titled
4	1.2.3	Audio Description or Media Alternative (Prerecorded)	17	2.4.3	Focus Order
5	1.3.1	Info and Relationships	18	2.4.4	Link Purpose (In Context)
6	1.3.2	Meaningful Sequence	19	3.1.1	Language of Page
7	1.3.3	Sensory Characteristics	20	3.2.1	On Focus
8	1.4.1	Use of Color	21	3.2.2	On Input
9	1.4.2	Audio Control	22	3.3.1	Error Identification
10	2.1.1	Keyboard	23	3.3.2	Labels or Instructions
11	2.1.2	No Keyboard Trap	24	4.1.1	Parsing
12	2.2.1	Timing Adjustable	25	4.1.2	Name, Role, Value
13	2.2.2	Pause, Stop, Hide			

The experts' evaluation process occurred between 1/May/2018 and 25/May/2018 and is based on WCAG 2.0 level A, as shown in Table 2. The experts were asked to identify the following measures when completing the task: time to complete, violating the success criteria, number of tabs and difficulty level. Note that if the evaluated content does not have a content that matches a success criterion, the success criteria are assumed to be satisfied as suggested by the W3C "Understanding Conformance"[8] document. For example, audio-only and video-only success criteria are not found in the evaluated content. As a result, these success criteria are recorded as satisfied.

In order to provide a perspective of the tested task, time is measured from the starting of the task until completing the task. The time needed to finish the evaluation of each task can help in measuring the task efficiency in web accessibility [47]. The common accessibility problems are investigated by identifying the violated success criteria and pointing out the reason for the problem. The number of keyboard tabs needed to complete a specific task is used as one of the indicators of the obstacles that hinder users who navigate the Web only using the keyboard from completing a task [18]. As the number of tabs increases, the degree of difficulty increases when trying to complete a given task. The difficulty level can take four different values: easy, medium, difficult and impossible.

4. Evaluation and Discussion

4.1 Automatic Tools

The following are the evaluation results for the automatic evaluation tools ACheker and Total Validator. In both tools, we choose WCAG 2.0 level A conformance. The URLs of the tested government E-services are shown in Table 3. Figure 1 shows the number of failed success criteria for the evaluated tasks in Kuwaiti government portal using the two tools. As the figure shows, Total Validator detects more errors of WCAG 2.0 level A in the portal homepage than ACheker. On the other hand, ACheker detects more errors of WCAG 2.0 level A for task 14 than Total Validator. Such observation verifies the benefit of using more than one automatic tool to evaluate accessibility to overcome drawbacks of a single tool use [38].

Table 4 shows the failed WCAG 2.0 (level A, AA and AAA) success criteria for Kuwaiti governmental websites using ACheker automatic tool.

As the table shows, the homepage and the tasks 14 and 16 have the highest number of failed success criteria. Note that the tasks 3, 6, 10, 11, 15 and 17 are not validated, either because we could not find a

[8] https://www.w3.org/WAI/WCAG21/Understanding/conformance#levels

Table 3. The E-service URLs.

No.	Ministry	Task	E-service URL
		E-gov. Portal Homepage	https://www.e.gov.kw/sites/kgoEnglish/Pages/HomePage.aspx
1	Public Authority for Civil Information	Inquiring about Civil ID status	https://www.e.gov.kw/sites/kgoEnglish/Pages/eServices/PACI/CivilIDStatus.aspx
2	Ministry of Interior	Violation Payment (Traffic & Immigration)	https://portal.acs.moi.gov.kw/wps/portal/violations
4	Ministry of Communication	Inquiring about phone bill and e-payment	https://www.e.gov.kw/sites/kgoenglish/Pages/eServices/MOC/BillsQuery.aspx#
5	Ministry of Electricity and Water	Electricity and water bills enquiry and e-payment	https://www.e.gov.kw/sites/kgoenglish/Pages/eServices/MEW/InquiryAboutBills.aspx
7	Ministry of Health	Request appointment for food checkup	https://www.e.gov.kw/sites/kgoEnglish/Pages/eServices/MOH/FoodCheckup.aspx
8	eNet Company	Mobile bill payment and recharge services	https://www.e.gov.kw/sites/kgoenglish/Pages/eServices/Enet/MobilePayments.aspx
9	Ministry of Education	Results of staff inward and outward transfer	https://www.e.gov.kw/sites/kgoenglish/Pages/eServices/MOE/InternalExternalShifiting.aspx
12	Public Authority for Civil Information	Multi-civil renewal and payments	https://www.e.gov.kw/sites/kgoEnglish/Pages/eServices/PACI/CivilIDRenewal.aspx
13	Ministry of Justice	Inquiring about travel ban	https://www.e.gov.kw/sites/kgoEnglish/Pages/eServices/MOJ/BanTravel.aspx
14	Ministry of Interior	Personal inquiry about MOI	https://www.e.gov.kw/sites/kgoenglish/Pages/eServices/MOI/PersonalInquiry.aspx
15	Public Authority for Civil Information	Civil ID fines e-payment	https://www.e.gov.kw/sites/kgoEnglish/Pages/eServices/PACI/FinesEPayment.aspx
16	Ministry of Interior	Inquiring about travel violations	https://www.e.gov.kw/sites/kgoenglish/Pages/eServices/MOI/EnviolationPlateNumber.aspx

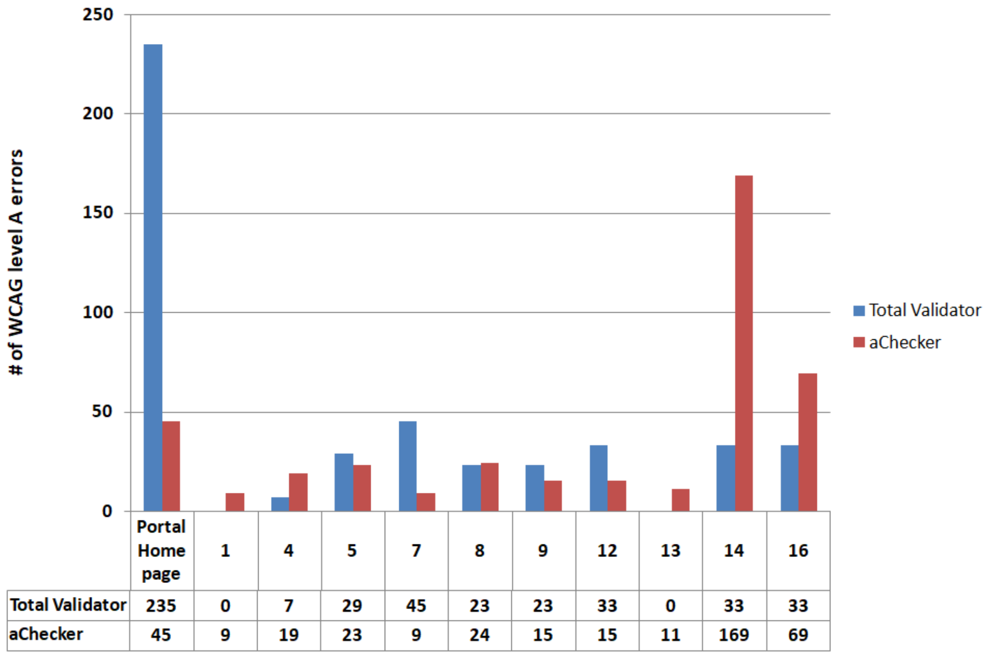


Figure 1. Failed WCAG 2.0 (level A) success criteria for Kuwaiti governmental websites using automatic tools.

Table 4. Failed WCAG (level A, AA and AAA) success criteria for Kuwaiti governmental websites using ACheker automatic tool.

			Known problems		
No.	Task Name	Ministry	A	AA	AAA
	E-gov. Portal Homepage		45	75	556
1	Inquiring about Civil ID status	Public Authority for Civil Information	9	24	327
4	Inquiring about phone bill and e-payment	Ministry of Communication	19	29	313
5	Electricity and water bills enquiry and e-payment	Ministry of Electricity and Water	23	37	331
7	Request appointment for food checkup	Ministry of Health	9	19	314
8	Mobile bill payment and recharge services	eNet Company	24	36	319
9	Results of staff inward and outward transfer	Ministry of Education	15	29	324
12	Multi-civil renewal and payments	Public Authority for Civil Information	15	26	332
13	Inquiring about travel ban	Ministry of Interior or Ministry of Justice	11	21	314
14	Personal inquiry about MOI	Ministry of Interior	169	179	346
16	Inquiring about travel violations	Ministry of Interior	69	79	326

link to the service or the access is prohibited. In order to investigate the usage of ARIA standards in the Kuwaiti government websites, the WAVE tool is utilized.

ARIA is usually used to enhance the Web content to be more accessible for screen reader users by placing landmarks, indicating the dynamically updated content and providing more semantic to the used Web widget through properties [45]. The results in Table 5 show that five (i.e., 50%) of the evaluated sites do not use ARIA standards.

4.2 HTML and CSS Evaluation

The conformance of the performed tasks to the HTML and CSS standards is verified using HTML and CSS validation services. Table 6 shows the total number of HTML and CSS validation errors in the evaluated Kuwaiti E-government services. The unavailable services are removed from the table. Note that the error "504 Gateway timeout" happened after providing the link to the online service and it seems that the requested page was not loaded successfully.

The total number of errors is calculated by validating HTML and CSS in each visited page when completing the defined task. We could not validate some of the tasks due to the inability to access the E-service; such errors are explained accordingly in the table.

As shown in Table 6, all the tasks have errors in HTML and CSS. The HTML validation errors range from 4 to 38 errors. On the other hand, the highest number of CSS validation errors is 129 and the lowest is zero. The Kuwaiti government portal homepage alone has 34 errors in both HTML and CSS. Such errors indicate a problem for the assistive technologies when they are utilized by people with disabilities to use the website.

4.3 Expert Review

The same set of tasks evaluated using automatic tools were evaluated by checking WCAG 2.0 level A

Table 5. The number of used ARIA in the evaluated Web pages. Highlighted rows are for unavailable.

No.	Task eService Name	Ministry	No. of Used ARIA
	E-gov. Portal Homepage		6
1	Inquiring about Civil ID status	Public Authority for Civil Information	7
2	Violation Payment (Traffic & Immigration)	Ministry of Interior	Error accessing page
3	Inquiring about lawsuits filled against you	Ministry of Justice	e-service unavailable
4	Inquiring about phone bill and e-payment	Ministry of Communication	0
5	Electricity and water bills enquiry and e-payment	Ministry of Electric and water	0
6	Renew work permit	Public authority of manpower	Server error 404 File or directory not found
7	Request appointment for food checkup	Ministry of Health	7
8	Mobile bill payment and recharge services	eNet company	0
9	Results of staff inward and outward transfer	Ministry of education	0
10	Inquiry into status of an application (altarasul system)	Civil Service Commission	No link to eservice found
11	Reserve a hall	Ministry of social affairs and labor	No link to e-service
12	Multi-civil renewal and payments	Public Authority for Civil Information	Error - page not publicly available
13	Inquiring about travel ban	Ministry of Interior	Error accessing page
14	Personal inquiry about MOI	Ministry of Interior	Error accessing page
15	Civil ID fines e-payment	Public Authority for Civil Information	2
16	Inquiring about travel violations	Ministry of Interior	0
17	Inquiring about arrest warrants	Ministry of Justice	Prohibited access

Table 6. Total number of HTML and CSS validation errors.

No.	Ministry	Task	# HTML errors	# CSS errors
		E-gov. Portal Homepage	34	34
1	Public Authority for Civil Information	Inquiring about Civil ID status	34	34
2	Ministry of Interior	Violations Payment (Traffic & Immigration)	I/O Error	504 Gateway timeout
4	Ministry of Communication	Inquiring about phone bill and e-payment	4	1
5	Ministry of Electricity and Water	Electricity and water bills enquiry and e-payment	38	0
7	Ministry of Health	Request appointment for food checkup	15	3
8	eNet Company	Mobile bill payment and recharge services	4	1
9	Ministry of Education	Results of staff inward and outward transfer	6	129
12	Public Authority for Civil Information	Multi-civil renewal and payments	I/O Error	I/O Error
13	Ministry of Justice	Inquiring about travel ban	I/O Error	504 Gateway timeout
14	Ministry of Interior	Personal inquiry about MOI	18	1
15	Public Authority for Civil Information	Civil ID fines e-payment	8	2
16	Ministry of Interior	Inquiring about travel violations	18	1

The experts were not able to complete the following tasks: 3, 6, 10, 11, 15 and 17 (see Table 1). This is because either the service was unavailable (3, 6 and 11; an example is shown in Figure 7) or the service was not found in the government portal (10, 15 and 17). Note that these tasks are removed from the results.

In each one of the tasks, the expert starts from the portal homepage and uses NVDA screen reader and keyboard only to complete the task. All the tasks require moving between different web pages. Some tasks need from the expert to fill text fields or forms. The task is considered completed when the expert receives the output from the service. The web browser Google Chrome is used to open the web pages.

Figure 2 presents the number of WCAG 2.0 failed success criteria for each task. The homepage of the government portal breaks three of the WCAG 2.0 level A success criteria, as shown in Figure 2, which are: "non-text content" as 15 images have no alternative text, "page titled" as the page title is not descriptive and "labels or instructions" as the text is spoken by the screen reader when the user reaches the search box on the Web page in a way that is not understandable (the screen reader reads the following text: "table L search e-payment civil employment insurance traffic electricity residency and immigration education service directory edit blank").

Figure 5 shows the percentage of WCAG 2.0 failed success criteria from all the evaluated tasks. Clearly, "non-text content" success criteria failed on all the tasks. In addition, "labels or instructions" success criteria failed for around 80% of the tasks. The following is a description of the reasons for failed success criteria:

- **On Focus**: the E-service page contains form elements and the focus is not inside the first element to insert the data (i.e., textbox). In addition, for all the E-services, the new page is opened in a new window when the user clicks the start E-service button.
- **Labels or Instructions**: the forms in the Web pages have no labels and the user cannot know what to enter. Also, when trying to submit the form information, inaccessible Captcha is required to continue the submission (Figure 6 is an example of this case).

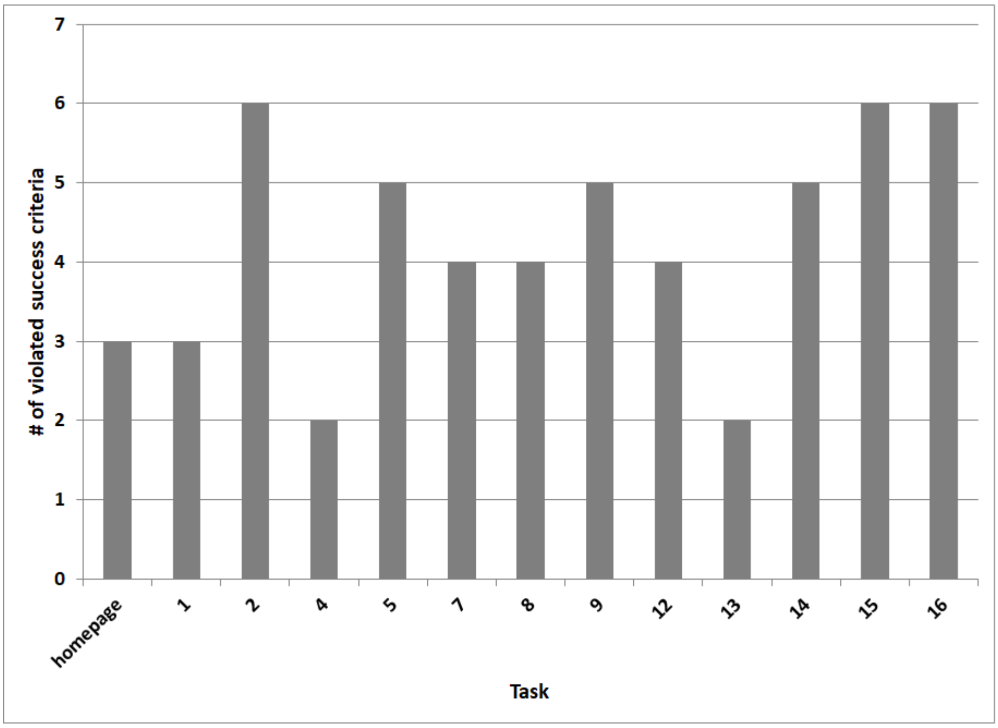
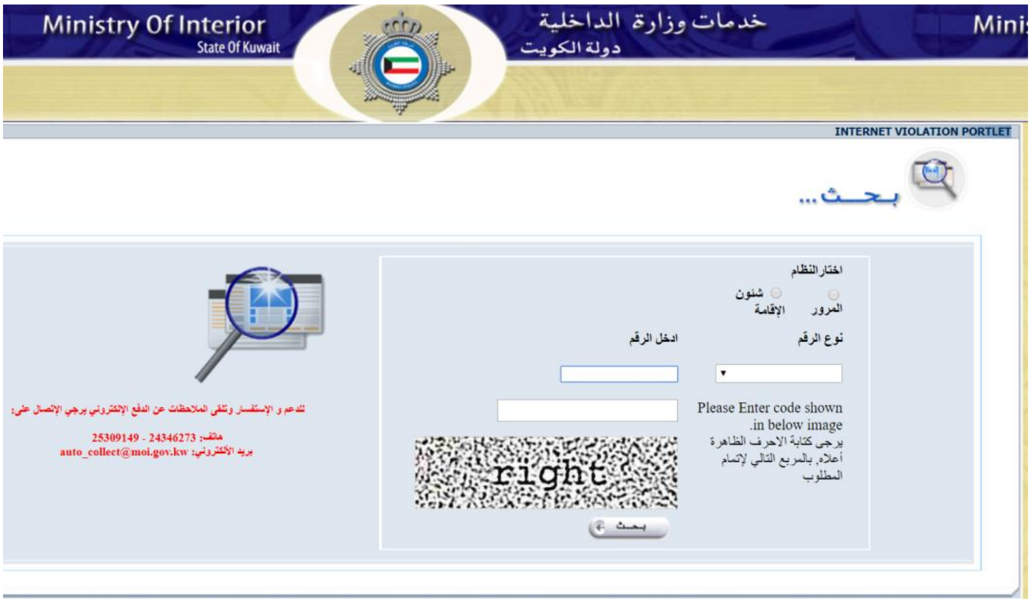


Figure 2. The number of violated guidelines per task using expert evaluation**.**

- **Meaningful Sequence**: when a link in the English section of the website takes the user to the Arabic homepage.



Please Enter code shown .in below image

يرجى كتابة الاحرف الظاهرة أعلاه, بالمربع التالي لإتمام المطلوب

للدعم و الإستفسار وتلقى الملاحظات عن الدفع الإلكتروني يرجي الإتصال على:

هاتف: 24346273 - 25309149

بريد الألكتروني: auto_collect@moi.gov.kw

Figure 3. A link in the English site for the task violation payment (traffic & immigration) takes the user to the Arabic homepage.

- **Link Purpose**: for all the evaluated tasks, the button to start the service has a general label (i.e., Start E-service).
- **Keyboard**: the "Frequently Used" service tab found in different pages when performing the tasks is not accessible using the keyboard.
- **Page Title**: a large number of the Web pages when completing the tasks have a title that is general and does not describe the current E-service page.

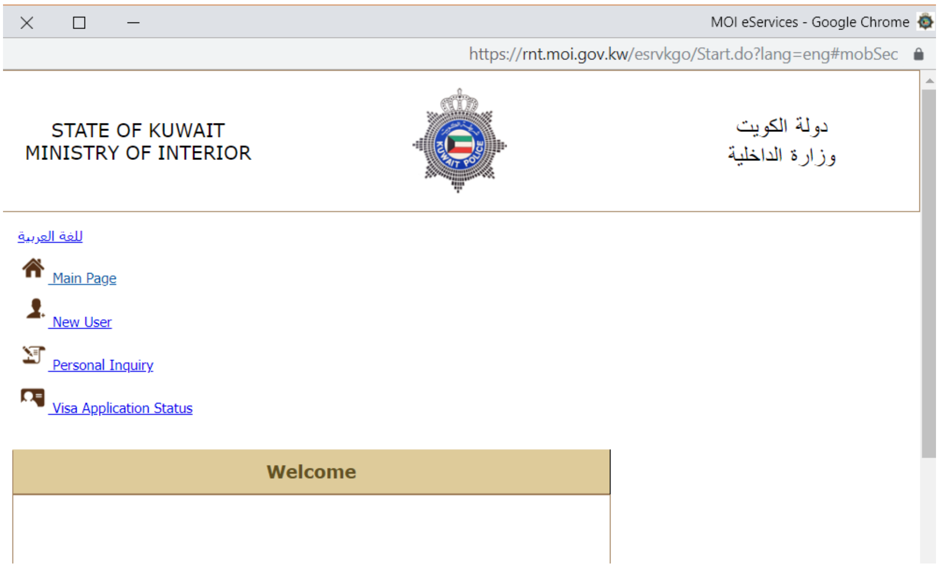


Figure 4. The E-service page for inquiring about travel violations with a general title which does not describe the E-service.

The experts' evaluation when performing different tasks based on time, the number of tabs and the difficulty when completing the tasks are shown in Table 7. The difficulty level in the Table takes the values: Medium (M), Difficult (D) and Impossible (I). Only four tasks out of the 17 are considered not impossible (i.e., 76% of the tasks are considered as impossible).

The task is rated as impossible in different situations which include: the service is not available (i.e., the service is down); the service is only keyboard accessible; while we are in the English site, it opens an Arabic page which cannot be read by screen reader; inaccessible Captcha needs to be used to submit the form and no labels available for the form presented to the user. Table 8 summarizes the tasks that were found impossible and their percentage from all the impossible tasks (see Table 6 for task name).

On the other hand, Figure 2 shows the number of failed success criteria on each of the validated tasks by the experts.

Table 7. Time, number of tabs and difficulty of each task (row in bold font is for impossible task).

Task No.	Time	Number of tabs	Difficulty
1	4:10	28	M
2	**NA**	**NA**	**I**
4	**NA**	**NA**	**I**
5	**NA**	**38**	**I**
7	**NA**	**56**	**I**
8	2:40	31	D
9	**NA**	**32**	**I**
12	2:40	32	M
13	**NA**	**NA**	**I**
14	4:15	24	D
15	**NA**	**NA**	**I**
16	**NA**	**37**	**I**

Table 8. Analysis of impossible tasks (see Table 6 for task name).

Reason	Tasks	% from impossible tasks
The service is down or not available	3, 6, 10, 11 and 17	38%
The service is only keyboard accessible	4 and 13	15%
The English site opens an Arabic Web page	2, 5, 7, 9, 15 and 16	46%
Inaccessible Captcha needs to be used to submit the form	2, 7, 15 and 16	30%
The form has no labels	2, 7, 8, 9, 15 and 16	46%

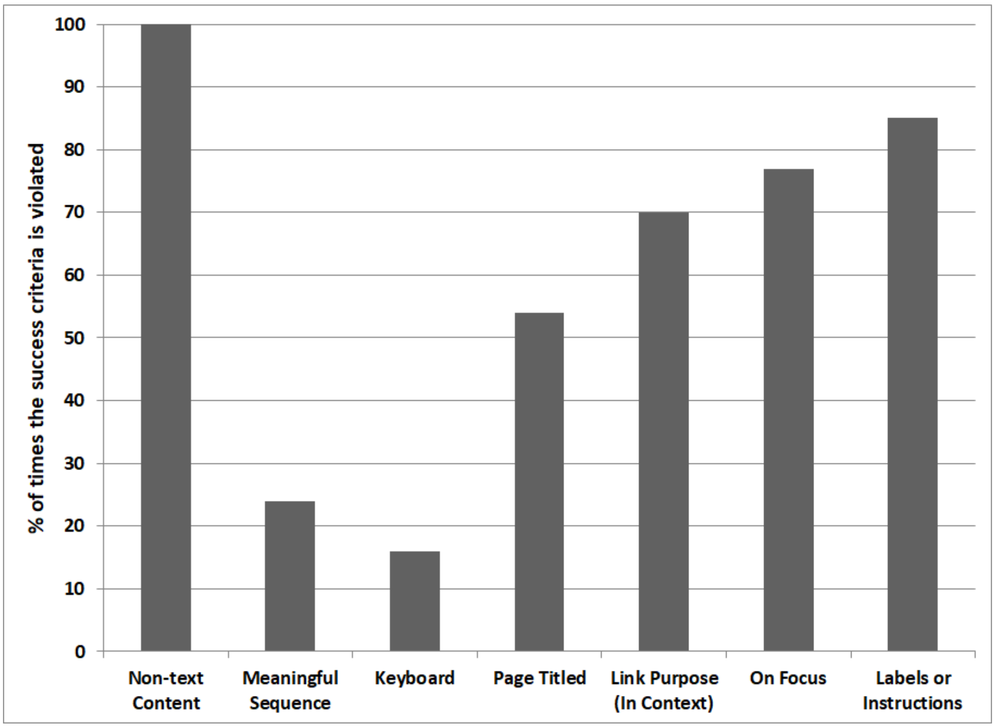
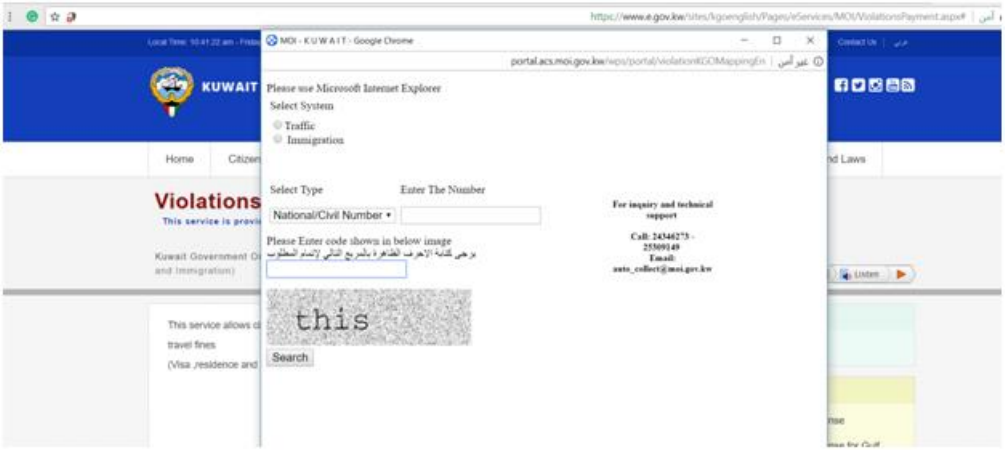


Figure 5. How many times the success criteria are violated.

4.4 Discussion

The findings suggest an urgent need to improve the accessibility of the Kuwaiti government portal. They reveal that a large number of the E-services are inaccessible. Violations of the WCAG 2.0 success criteria include: non-text content, link purpose (in context), on focus and labels or instructions.

Several of the basic citizen E-services tested were impossible to use. As shown in Table 6, 46% of the target E-services were down or not available and 46% of the E-services contain forms with no labels (a major reason for the inaccessibility of the E-services).


https://www.e.gov.kw/sites/kgoenglish/Pages/eServices/MOI/ViolationsPayment.aspx#
MOI - K U W A I T - Google Chrome
portal.acs.moi.gov.kw/wps/portal/violationKGOMappingEn
KUWAIT
Home
Citizen
Violations
d Laws
Listen

Please use Microsoft Internet Explorer

Select System

○ Traffic

○ Immigration

Select Type | Enter The Number

National/Civil Number ▾

Please Enter code shown in below image

يرجى كتابة الأحرف الظاهرة بالمربع التالي لإتمام المطلوب

this

Search

For inquiry and technical support

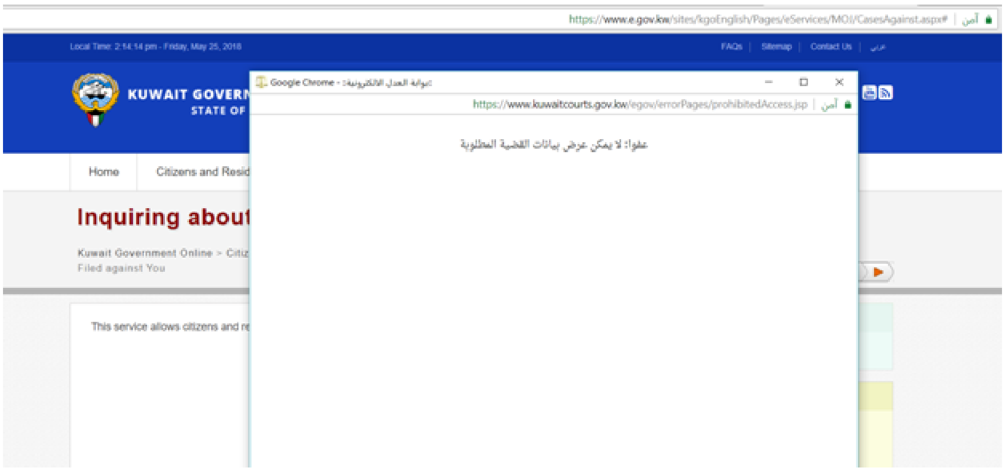
Call: 24346273 - 25309149

Email: auto_collect@moi.gov.kw

Figure 6. Violation payment (traffic & immigration) service asking user to enter inaccessible Captcha.

In addition, some of the English pages open as Arabic pages with no way to reach the English text, which is excluding a large number of citizens in Kuwait who speak only English. These violations have relatively simple solutions, but the knowledge and awareness of how to cater for people with disabilities in Kuwait remain relatively low.

Note that the following tasks are considered reachable by the automatic tools as the first page of the service is tested (1, 4, 5, 7, 8, 9, 12, 13, 14 and 16). On the other hand, the following tasks: (4, 5, 7, 8, 9, 13 and 16) are considered impossible when tested by the experts, as shown in Table 6.


https://www.e.gov.kw/sites/kgoEnglish/Pages/eServices/MOJ/CasesAgainst.aspx#
Local Time: 2:14:14 pm - Friday, May 25, 2018
FAQs
Sitemap
Contact Us
عربي
KUWAIT GOVERN
STATE OF
Google Chrome - بوابة العدل الالكترونية
https://www.kuwaitcourts.gov.kw/egov/errorPages/prohibitedAccess.jsp
Home
Citizens and Resid
Inquiring about
Kuwait Government Online > Citiz
Filed against You
This service allows citizens and re

عفواً: لا يمكن عرض بيانات القضية المطلوبة

Figure 7. Inquiring about lawsuits service is not available.

4.5 Policy, Legislation and Awareness

The Kuwait National Framework for Digital Accessibility suggests that for the Web and documents to be accessible, they should conform to the basic criteria of the Web Content Accessibility Guidelines (WCAG 2.0). Following the official announcement in 2018 [10], the government failed to offer a plan for implementation at the national and organizational levels.

To be able to effectively assess the conformance of the government portal and sustain it, its host CAIT should be required to establish an accessibility policy of its own, set conformance milestones and monitor and review the website on a regular basis. Some entities in Kuwait have begun to implement elements of the WCAG 2.0 and the Framework by enhancing the user experience with the goal of growing the margin of profit.

Kuwait Vision 2035 in accessibility is about how to make Kuwait accessible to everyone, including people with disabilities, in different fields, such as physical and digital ones. Physical accessibility is enhanced by using universal design to be applied to make the environment more accessible. Digital accessibility is achieved by using Kuwait National Framework for Digital Accessibility to be implemented on the digital technology and not only on the Web.

In addition, the national accessibility framework is merely a policy, not a law. The next step should be to pass a law through the parliament, so that the government can setup a legal framework that will hold people accountable if it is violated. Without policies and laws or fear of prosecution, there are no incentives for the government or private entities to begin enhancing their websites, applications and services to engage people with disabilities.

Finally, E-government services should be available for all the citizens in the country no matter what their language, ability or age is. With the impending adoption of cloud-based services, there is a need to re-structure the E-government landscape to facilitate user tasks, which will help elicit more e-service usage. This puts a strain on Web developers who should be aware of the national framework and trained on accessibility standards.

Developers should make accessibility a core part of all their development projects, particularly the E-government services.

5. Conclusion and Future Work

In this paper, the accessibility of Kuwaiti E-government services is evaluated. We tested the accessibility of 17 of the top used E-services in 2018 against WCAG 2.0 level A. The study applies different technical dimensions to investigate accessibility problems using automatic tools and experts' manual review.

The overall results show that most of the evaluated E-government websites lack accessibility. Unfortunately, thirteen out of seventeen (i.e., 76%) of the evaluated E-services are impossible to use. The most commonly failed accessibility success criteria are: "non-text content" and "labels or instructions".

Furthermore, Web accessibility guidelines are not mentioned in the government portal or by the public authority of the disabled which is responsible for people with disabilities in Kuwait. There is a need to further develop appropriate policies and laws and set a national level plan to enforce the adoption of the national accessibility guidelines and WCAG standards for better inclusion of all residents and citizens in Kuwait.

In most of the studies that evaluate the accessibility of E-government, only the homepage is used to check whether or not it complies with WCAG [46]. In this study, a task-based technique is used which navigates different pages to evaluate the accessibility of the provided E-service. This helps in checking the accessibility of the provided E-service in a situation similar to what the user encounters in the real life. Furthermore, this enables the checking of the ease of navigation between the pages to complete the service.

Future assessments may involve users with disabilities in the testing of the site, as they can provide a more realistic assessment of the website's accessibility. In addition, performance indicators for Web accessibility need to be used to watch the country improvement in terms of Web accessibility and an analysis of Web developer awareness of accessibility standards is needed.

References

[1] WAI, Introduction to Web accessibility, [Online], Available: https://www.w3.org/WAI/.

[2] UN, United Nations E-government Survey, [Online], Available: https://publicadministration.un.org/egovkb/en-us/reports/un-e-government-survey-2016.

[3] T. Nam, "Determining the Type of E-government Use," Government Information Quarterly, vol. 31, no. 2, pp. 211-220, 2014.

[4] WorldBank, The World Bank's Governance Global, [Online], Available: http://documents.worldbank.org/curated/en/833041539871513644/pdf/131020-WP-P163620-WorldBankGlobalReport-PUBLIC.pdf

[5] The Supreme Council for Planning, Development, Kuwait National Development Plan, [Online], Available: https://kif.kdipa.gov.kw/wp-content/uploads/khalid-mahdi-english.pdf, 2016.

[6] M. Saleem, Web Accessibility Compliance for e-Government Websites in the Gulf Region, Master's Thesis, Edith Cowan University, Australia, 2016.

[7] M. Batusic, A. Gaal, J. Klaus and M. O'Grady, "An IT Training Programme for Blind Computer Users: Presentation and Discussion of Didactic and Teletutorial Implications," Computers Helping People with Special Needs, Springer, Berlin-Heidelberg, pp. 1306-1312, 2006.

[8] United Nations, Convention on the Rights of Persons with Disabilities, [Online], Available: https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html, 2006.

[9] Kuwait Law No. 8 for Persons with Disabilities of 2010, [Online], Available: https://www.ilo.org/dyn/natlex/docs/ELECTRONIC/89501/102841/F-1202766234/KALD20110318.pdf

[10] "Kuwait Digital Accessibility Framework," Workshop UNDP in Kuwait, [Online], Available: http://www.kw.undp.org/content/kuwait/en/home/presscenter/articles/_kuwait-digital-accessibility-framework-workshop.html.

[11] Achieving Kuwait 2035 Vision Towards Persons with Disability, [Online], Available: http://www.kw.undp.org/content/kuwait/en/home/operations/projects/human_development/achieving-kuwait-2035-vision-towards-persons-with-disability.html.

[12] B. Leporini and F. Paternò, "Increasing Usability When Interacting through Screen Readers," Universal Access in the Information Society, vol. 3, no. 1, pp. 57-70, 2004.

[13] WCAG 2.0, Web Content Accessibility Guidelines, [Online], Available: https://www.w3.org/TR/WCAG20/

[14] Y. Akgul, "Website Accessibility, Quality and Vulnerability Assessment: A Survey of Government Websites in the Turkish Republic," Jou. of Inf. Sys. Eng. & Manag., vol. 4, no. 1, pp. 1-13, 2016.

[15] W. Yaokumah, S. Brown and R. Amponsah, "Accessibility, Quality and Performance of Government Portals and Ministry Websites: A View Using Diagnostic Tools," Annual Global Online Conference on Information and Computer Technology (GOCICT), pp. 46-50, 2015.

[16] H. Wanyonyi Kituyi and J. Waweru, "Evaluation of the Accessibility of Kenya e-Government Websites in the Nairobi Central Business District," European Jou. of Technology, vol. 1, no. 1, pp. 36-55, 2016.

[17] A. Al-Faries, H. S. Al-Khalifa, M. S. Al-Razgan and M. Al-Duwais, Evaluating the Accessibility and Usability of Top Saudi e-Government Services," Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance (ICEGOV '13), ACM, NY, USA, pp. 60-63, 2013.

[18] I. A. Doush, A. Bany-Mohammed, E. Ali and M. A. Al-Betar, "Towards a More Accessible e-Government in Jordan: An Evaluation Study of Visually Impaired Users and Web Developers," Behaviour & IT, vol. 32, no. 3, pp. 273-293, 2013.

[19] I. A. Doush and I. Alhami, "Evaluating the Accessibility of Computer Laboratories, Libraries and Websites in Jordanian Universities and Colleges," International Journal of Information Systems and Social Change (IJISSC), vol. 9, no. 2, pp. 44-60, 2018.

[20] B. Al Mourad and F. Kamoun, "Accessibility Evaluation of Dubai e-Government Websites: Findings and Implications," Journal of E-Government Studies and Best Practices, vol. 2013, Article ID 978647, DOI: 10.5171/2013, 2013.

[21] F. Kamoun and M. B. Almourad, "Accessibility As an Integral Factor in e-Government Web Site Evaluation: The Case of Dubai e-Government," Information Technology and People, vol. 27, no. 2, pp. 208-228, 2014.

[22] H. M. Park, "The Web Accessibility Crisis of the Korea's Electronic Government: Fatal Consequences of the Digital Signature Law and Public Key Certificate," Proc. of the 45th Hawaii International Conf.

on System Sciences, pp. 2319-2328, 2012.

[23] M. R. Patr, A. R. Dash and P. K. Mishra, "Accessibility Analysis of Government Web Portals of Asian Countries," Proceedings of the 8th Int. Conf. on Theory and Practice of Electronic Governance (ICEGOV '14), ACM, New York, NY, USA, pp. 383-386. doi:10.1145/2691195.2691253, 2014.

[24] Y. M. Tashtoush, A. F. Darabseh and H. N. Al-Sarhan, "The Arabian e-Government Websites Accessibility: A Case Study," Proc. of the 7th International Conference on Information and Communication Systems (ICICS), pp. 276-281, 2016.

[25] A. Ismail and K. S. Kuppusamy, "Accessibility of Indian Universities’ Homepages: An Exploratory Study," Journal of King Saud Uni.-Computer and Information Sci., vol. 30, no. 2, pp. 268-278, 2018.

[26] H. S. Al-Khalifa, "The Accessibility of Saudi Arabia Government Websites: An Exploratory Study," Universal Access in the Information Society, vol. 11, no. 2, pp. 201-210, 2012.

[27] H. S. Al-Khalifa, I. Baazeem and R. Alamer, "Revisiting the Accessibility of Saudi Arabia Government Websites," Universal Access in the Information Society, vol. 16, no. 4, pp. 1027-1039, 2017.

[28] M. Vigo and G. Brajnik, "Automatic Web Accessibility Metrics: Where We Are and Where We Can Go," Interacting with Computers, vol. 23, no. 2, pp. 137-155, 2011.

[29] M. Akram and R. B. Sulaiman, "A Systematic Literature Review to Determine the Web Accessibility Issues in Saudi Arabian University and Government Websites for Disabled People," International Journal of Advanced Computer Science and Applications, vol. 8, no.6, pp. 321-329, 2017.

[30] Y. AkgUL and K. Vatansever, "Web Accessibility Evaluation of Government Websites for People with Disabilities in Turkey," Journal of Advanced Management Science, vol. 4, no. 3, pp. 201-210, 2016.

[31] A. Olalere and J. Lazar, "Accessibility of U.S. Federal Government Homepages: Section 508 Compliance and Site Accessibility Statements," Government Information Quarterly, vol. 28, no. 3, pp. 303 – 309, 2011.

[32] IT Accessibility Laws and Policies, [Online], Available: https://www.section508.gov/manage/laws-and-policies

[33] P. T. Jaeger, "Assessing Section 508 Compliance on Federal e-Government Websites: A Multi-method, User-centred Evaluation of Accessibility for Persons with Disabilities," Government Information Quarterly, vol. 23, no. 2, pp. 169 – 190, 2006.

[34] M. H. A. Latif and M. N. Masrek, "Accessibility Evaluation on Malaysian e-Government Websites," Article ID 935272, pp. 1-11, 2010.

[35] M. Saleem, "Arabic Web Accessibility Guidelines: Understanding and Use by Web Developers in Kuwait," Proc. of the Internet of Accessible Things (W4A '18), ACM, NY, USA, pp. 1-25, 2018.

[36] A. Alayed, A Framework and Checklist for Localized Web Content Accessibility Guidelines for Arabic University Websites in Saudi Arabia, Ph.D. Thesis, University of Southampton, UK, 2018.

[37] Kuwait E-Government Portal Statistics 2018, [Online], Available: https://www.e.gov.kw/sites/kgoArabic/Pages/InfoPages/Statistics.

[38] M. Vigo, J. Brown and V. Conway, "Benchmarking Web Accessibility Evaluation Tools: Measuring the Harm of Sole Reliance on Automated Tests," Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, ACM, pp. 1-10, 2013.

[39] I. S. Baazeem and H. S. Al-Khalifa, "Advancements in Web Accessibility Evaluation Methods: How Far Are We?," Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services, ACM, pp. 90-95, 2015.

[40] AChecker, ATRC Web Accessibility Checker [Online], Available: http://achecker.ca/checker/index.php, [Accessed: 19-05-2018].

[41] TotalValidator, Total Validator, [Online], Available: https://www.totalvalidator.com, [Accessed: 19-05-2018].

[42] WAVE, WAVE Web Accessibility Tool, [Online], Available: http://wave.Webaim.org, [Accessed: 19-05-2018].

[43] M. K. Baowaly and M. Bhuiyan, "Accessibility Analysis and Evaluation of Bangladesh Government Websites," Proc. of IEEE International Conference on Informatics, Electronics & Vision (ICIEV), pp. 46-51, 2012.

[44] J. Grantham, E. Grantham and D. Powers, "Website Accessibility: An Australian View," Proceedings of the 13th Australasian User Interface Conf.-Volume 126, Australian Comp. Soci., Inc., pp. 21-28, 2012.

[45] I. A. Doush, F. Alkhateeb, E. A. Maghayreh and M. A. Al-Betar, "The Design of RIA Accessibility Evaluation Tool," Advances in Engineering Software, vol. 57, no. 2013, pp. 1-7, 2013.

[46] L. Moreno, P. Martinez, J. Muguerza and J. Abascal, "Support Resource Based on Standards for Accessible e-Government Transactional Services," Computer Standards & Interfaces, vol. 58, pp. 146-157, 2018.

[47] G. Brajnik, "A Comparative Test of Web Accessibility Evaluation Methods," Proceedings of the 10th International ACM SIGACCESS Conference on Comp. and Accessibility, ACM, pp. 113-120, 2008.

[48] J. P. Bigham, C. M. Prince and R. E. Ladner, "WebAnywhere: A Screen Reader on-the-go," Proceedings of the International Cross-disciplinary Conference on Web Accessibility (W4A), ACM, pp. 73-82, 2008.

[49] S. Harper, Y. Yesilada and T. Chen, "Mobile Device Impairment… Similar Problems, Similar Solutions?," Behaviour & Information Technology, vol. 30, no. 5, pp. 673-690, 2011.

[50] S. G. Abduganiev, "Towards Automated Web Accessibility Evaluation: A Comparative Study," Int. J. Inf. Technol. Comput. Sci. (IJITCS), vol. 9, no. 9, pp. 18-44, 2017.

[51] W3C, Understanding WCAG 2.0, [Online], Available: https://www.w3.org/TR/UNDERSTANDING-WCAG20/conformance.html.

[52] G. M. Alsalem and I. A. Doush, "Access Education: What Is Needed to Have Accessible Higher Education for Students with Disabilities in Jordan?," International Journal of Special Education, vol. 33, no. 3, pp. 541-561, 2018.

[53] E. Ali, I. Abu Doush, G., Alsalem and W. Alrashdan, "Evaluating the Web Accessibility of University Online Registration System: Case Study on Jordan," International Journal of Advanced Science and Technology, vol. 13, 2019.

ملخص البحث:

تستخدم الشبكة العنكبوتية من أجل توفير أنواع مختلفة من الخدمات للناس. وفي هذا الشأن، يجب أخذ احتياجات الأشخاص ذوي الإعاقة في الاعتبار عند تطوير الحكومة الإلكترونية. ويهدف هذا العمل الى تقييم قابلية الوصول الى المواقع الإلكترونية والعقبات التي تعترض الأشخاص ذوي الإعاقة في الوصول الى تلك المواقع الخاصة بالحكومة الإلكترونية في الكويت، وذلك في محاولة لتحديد المشكلات وزيادة الوعي لدى الحكومة بتلك المشكلات وتحسين شمول الأشخاص ذوي الإعاقة بالخدمات المقدمة. ولتقييم قابلية الوصول الى خدمات الحكومة الإلكترونية في الكويت، تم تطبيق التقييم الآلي وتقييم الخبراء على أبرز سبع عشرة خدمة إلكترونية مستخدمة عام 2018 في الكويت. وبينت النتائج التي تم الحصول عليها أن 13 من الخدمات هي غير ممكنة الاستخدام؛ الأمر الذي ينم عن ضعف شديد في التقيد بإرشادات الوصول الى محتوى الشبكة العنكبوتية (WCAG 2.0 المستوى A لغالبية المواقع الخاضعة للتقييم. وأظهرت الدراسة أهمية اتباع نهج قائم على المهام عند تقييم قابلية الوصول الى المواقع الإلكترونية، نظراً لأن الإبحار بين الصفحات المختلفة بعد الانتهاء من كل مهمة يمكن أن يساعد في اكتشاف قضايا أخرى تتعلق بقابلية الوصول.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 05, No. 03, December 2019.

AUTOMATED ARABIC ESSAY GRADING SYSTEM BASED ON F-SCORE AND ARABIC WORDNET

Saeda A. Al Awaida[1], Bassam Al-Shargabi[1] and Thamer Al-Rousan[2]

(Received: 10-Jun.-2019, Revised: 29-Jul.-2019 and 20-Aug.-2019, Accepted: 24-Aug.-2019)

ABSTRACT

An Automated Essay Grading (AEG) system is designed to be used in universities, companies and schools, which depends on Artificial Intelligence and Natural Language Processing technologies; as it has the capability to improve the grading system in terms of overcoming cost, time and teacher effort while correcting the students' essay questions and papers. The AEG system widespread use is due to its cost, accountability, standards and technology; as that leads to the system being used and applied for multiple languages, such as English and French, among others. On the other hand, limited research has been conducted to automate Arabic essay grading. Therefore, this paper introduces an Arabic AEG system. In this paper, we propose a model for Arabic essay grading based on F-score to extract features from student answers and model answers along with the use of the Arabic WordNet (AWN) as a valuable knowledge-based method for semantic similarity. The purpose of using the AWN is to find all related words from student answers to give the answer a student score. Students will not be subject to injustice in terms of their marks in cases when they do not write the exact model answer, which subsequently leads to an improvement of the Arabic AEG system to match human grading. The proposed model is evaluated using Arabic essay dataset and the result shows that our proposed model produces a result which matches human grading.

KEYWORDS

Automated Essay Grading, Support Vector Machine, Arabic WordNet, Cosine Similarity, Natural Language Processing.

1. INTRODUCTION

Automated Essay Grading (AEG) techniques are used in grading student essays without the direct participation of individuals, where an AEG system can automatically evaluate and produce a score or grade for a written essay to tackle time, reliability and cost issues. AEG systems are motivated by the need to develop solutions to assist teachers in grading essays in an efficient and effective manner.

AEG systems keep on drawing the interest of government-funded schools, colleges, testing organizations, specialists and instructors. Numerous studies have been conducted to examine the accuracy and precision of AEG systems. Furthermore, there were several studies conducted on AEG systems which revealed high matching rates between AEG system scores and human scores with various AEG systems [1]. The idea of having compelling approaches and techniques to score essays of students is to liberate instructors of the burden of perusing and hand-scoring possibly hundreds of essays and papers.

Moreover, test publishers would most likely score essays and papers for a cheaper expense and possibly give higher-quality assigned scores by using the computer's special capabilities to improve AEG systems to achieve more accurate results compared to traditional scoring. AEG system mechanism contains many stages: collecting the student texts in text corpus form inputted into the AEG software. Firstly, the AEG system pre-processes the texts to make them useful for further processing and analysis. The basic pre-process technique includes stripping the texts of white spaces, removing certain characters such as punctuation, removing any character from other languages and splitting the text sequence into pieces, referred to as tokens. Other methods employed in the pre-processing will be illustrated in more detail in the next sections.

The second stage typically involves feature extraction, which is concerned with mapping the text sequence into a vector of measurable quantities, the most common examples and the frequency of each unique word in the text. It is considered as the most difficult part of the construction of an AEG system

1. S. A. Al Awaida and B. Al- Shargabi are with Middle East University, Amman, Jordan. E-mails: `Saeda0awaida@gmail.com` and `bshargabi@meu.edu.jo`
2. T. Al Rousan is with Isra University, Amman, Jordan. E-mail: `thamer.rousan@iu.edu.jo`

and it is challenging for humans to take into account all the factors affecting the grade. Furthermore, the effectiveness of the AEG system is constrained by the chosen features [2].

Many techniques were exploited to automate essay grading; techniques within the field of natural language processing, latent semantic analysis and machine learning [3]. Automated Arabic essay grading is still at its beginnings with limited research conducted in this field. In this direction, we propose in this paper a model to automate Arabic essay grading based on F-score to extract features from student answers and model answers along with the use of Arabic WordNet (AWN), which is a valuable system for semantic similarity measures and text similarity algorithms. The use of AWN is useful to find all related words from the student's answers, which match the meaning of such words in the model answer; in order to facilitate grading the students' essays. Students are not subject to injustice regarding their marks in cases when they do not write the same exact model answer, which subsequently leads to the improvement of the AEG system to match human grading.

2. LITERATURE REVIEW

This section presents an overview of the concepts and main topics of Arabic automated essay grading, which includes, F-score, AWN and the most recent related work.

2.1 F-score

Support vector machine is represented by sparse vectors under the vector space, where each word in the vocabulary is mapped into one coordinate axis. This is used on data to train a linear classifier which is characterized by the normal to the hyper-plane dividing positive and negative instances.

We apply feature selection aiming to pre-define the number of the highest scoring features to be included in a classifier by using the F-score technique. F-score is a feature selection technique in SVM. F-score measures the distinction between two classes (positive and negative), where each feature is assigned to a value computed as in [4]. If that value of F-score for the feature is bigger than the mean value of all F-scores in order for the feature to be added to feature space, the feature will not be considered for the feature space.

F-score is used in the proposed model to decide or select the feature, which affects the score of the student answer by determining the positive and negative classes according to a related or non-related answer, where the related answer (positive) takes it, while the others (negative) ignore it. F-score is good for feature selection, where it solicits each feature separately based on its score over the Fisher criterion, which prompts an optimal subset of features, especially when the features are extracted from text like essays to redundant features.

2.2 Text Similarity Algorithms

Many text similarity approaches have been used to develop automated essay grading systems [5]-[6]. There are three major approaches for text similarity: string-based similarities, corpus-based similarities and knowledge-based similarities, in addition to a sample of combinations of all of them. String-based similarities are also divided into two types; character-based and term-based, where these approaches measure similarity by counting the number of different characters in two sequences.

Corpus-based similarities are similarity measures between words based on information collected from a huge amount of texts which are mainly used for language research. The knowledge-based similarity is a semantic similarity measure, which relies on determining the ratio of similarity between texts using information collected from the semantic network. Moreover, some of these approaches are combined together to find optimal performance in terms of accuracy.

2.3 Arabic WordNet

Arabic WordNet is a valuable knowledge-based tool for several semantic similarity measures. It was created in 2006 and expanded in 2016 [7]. AWN is a lexical database for the Arabic language which is concerned with the meaning of words, rather than forms, where words are semantically similar. Moreover, its lexical resources contain not only words of the targeted language, but also synsets and semantic relations between words, such as synonymy, meronymy and antonymy; as synsets are groups of words that can substitute other words in a sentence without changing its general meaning [7]. In this paper, we used Arabic WordNet to find all the related words from the student answers to give the student answer a score. Students are not subject to injustice regarding their marks when they do not write the same exact model answer.

2.4 Related Work

The latest and most recent related work is presented in this subsection, where many different approaches and systems were used to automate Arabic essay grading. An Arabic essay system called 'Abbir ' is presented in [8], where latent semantic analysis was used with some features, such as word stemming, spelling mistakes, proportion of spelling mistakes and word frequency, which revealed after different experiments that the performance is very close to human rating.

An automated assessor proposed in [7,9] for Arabic free text answer is based on LSA, which relies on replacing synonyms for each of the selected features to produce a matrix which is better than the traditional form of LSA matrix. The authors used the cosine similarity metrics to measure the similarity degree between the questions' model answers and the student answers. Accordingly, the score is given based on the higher ratio of similarity to set a score for the current essay based on model answer degrees.

A modified LSA is proposed in [9] for automatic essay scoring using Arabic essay answers, where a combined method of syntactic feature and LSA is based on bag-of-words. Afterwards, pre-processing creates a matrix, then applies cosine to define similarity. Results showed that syntactic feature improves accuracy. In this paper, we use AWN to apply the meaning features.

Moreover, a hybrid method employing LSA and rhetorical structure theory for automated Arabic essay scoring is proposed in [10], where the essay is semantically analyzed using LSA along with assessing essay writing style cohesiveness. The essay score of this approach was assigned based on the cohesion of the essay which represents 50% of the score, while 40% of the score is based on the writing style and the rest is given based on the spelling mistakes.

In [11], the authors suggested a web-based system which relies on using the Vector Space Model (VSM) to automate essay grading written in Arabic language. The system relies on two main processes; the first process extracts the features from essays, then applies SVM to find out the similarities between the essays written by the teachers and the ones written by the students, after converting each essay to vector space. The system then uses VS to match terms in the document after which cosine similarity is applied to find the score of student's answer. In this paper, SVM is used to extract features from answers. Another automated essay scoring system proposed in [12] also relies on using SVM, as it first extracts numerical features vector from the text data of essays using support vector machine classifier, then constructs a predictive model with extracted features and solves the multi-classification problem into multiple binary classifications to find the score between pairs of classes.

An approach using multiple classifiers, such as SVM, K-NN and Naïve Bayes in classifying Arabic language text documentation and comparing between those classifiers is used in [13]. The researchers used a dataset from Aljazeera news website and Al-Hayat website according to certain measures (recall, precision and F1), where the result suggests that the SVM classifier significantly outperforms other classifiers in high dimensional feature spaces. Accordingly, as the results in [13]-[14] indicate, F-score employed to extract features was the main reason of the model's high accuracy, which is why we used it in our proposed model.

A survey of similarity methods which focuses on challenges facing Arabic texts is employed in [15]. Three types of similarities were surveyed; lexical similarity based on character and statement similarities, semantic similarity and a hybrid similarity which combines both lexical and semantic similarities. The approach concluded that the cosine similarity metric produces an efficient performance when used in many Arabic essay grading systems, compared with other lexical measurements.

Moreover, a system based on the comparison of different text similarity algorithms for Arabic essay grading, such as string algorithm and corpus algorithm, is presented in [16]. The researchers applied multiple similarity measures to find an efficient way for essay grading. The N-gram approach is used in their system, as they relied on N-gram approach simplicity, which produces a reliable outcome when it comes to noisy data, such as grammatical errors or spelling mistakes, compared with the word-based approach.

A short answer system, based on translating student answers written in Arabic language into English language, is presented in [5] to tackle the challenges of the Arabic text in [6]. Accordingly, some problems existed during the translation process, such as a word in Arabic not in the same context structure and semantic is translated. Afterwards, the system applies multiple similarity measures and combines them to define the score of the tested student's answer. In this paper, we directly apply the similarity measure after extracting the feature without translation.

A system to automate essay scoring for online exams in Arabic language, based on evaluating the effects of stemming techniques, is applied in [17]. Heavy stemming and easy (light) stemming along with Levenshtein similarity measure are applied to the question in order to check the effectiveness of both techniques. As light stemming halts the elimination of prefixes and suffixes, without the ability to recognize the root of the word, heavy stemming is a root-based stemming which relies on eliminating prefixes and suffixes to get the actual root of a word. After finding the stemming word, the Levenshtein similarity measure is applied by giving each word a weight, then defining the distance between every two words to find the score.

3. PROPOSED MODEL

In this section, we present the proposed Arabic automated essay grading model, as illustrated in Figure 1, which is based on F-score, AWN and text similarity; to enhance the accuracy of the grading of essay exams. We developed a dataset (corpus) which is created to test the model. The proposed model consists of many phases, such as pre-processing, Arabic WordNet, feature extraction, using F-score and finally, applying the cosine similarity measure to determine the score of the student's answer; based on its cosine similarity degree and the model answers.

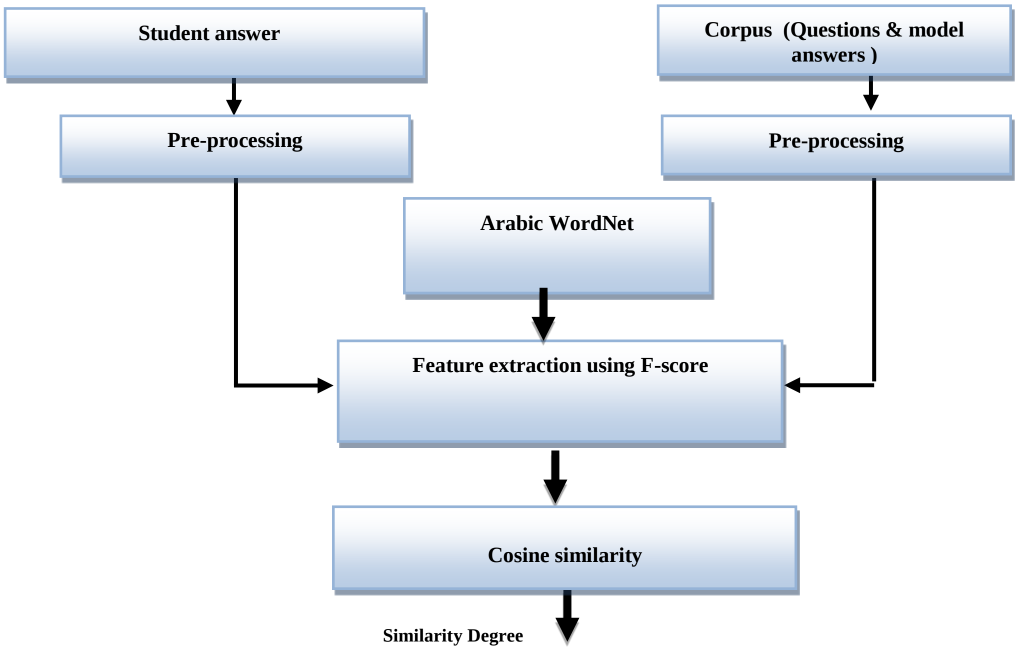


Figure 1. Proposed model.

3.1 Pre-processing

As seen in Figure 2, the pre-processing's first step is tokenization and normalization, which splits strings of student answer and model answer into smaller pieces and processes the transformation of the characters and the words into a single form. The second step is stop-word removal, where stop-words can be defined as words that do not have any significant meaning, or any word which does not have any importance and meaning in terms of finding the text classification; as these words are removed from the text. After converting the input Arabic text into a list of tokens, they are inputted to the next stage which is the stop-word removal, as they will be listed in the dictionary to remove them from the tokens output. The third stage is the stemming and lemmatizing process , which is a procedure's function for retrieving the word to its basic root, by processing the removal of all of the word prefixes, suffixes and infixes. Lemmatization is closely related to stemming which extracts the base root of words. It creates an actual dictionary for words.

Example of the pre-processing step is as follows: An example of a student’s answer:

مجموعة من الحواسيب والأدوات والمعدات ترتبط فيما بينها بوساطة خطوط اتصال

The result after the tokenization process:

""مجموعة" " من " " الحواسيب " "الأدوات" " ترتبط " " فيما " " بينها" " "المعدات" " بوساطة" " خطوط" " اتصال "

Stop word removal result:

"مجموعة" "الحواسيب " "المعدات" "الأدوات" " ترتبط " "بوساطة" "خطوط" "اتصال"

Different types of stemmers are used for Arabic text; in the proposed model, we used ISRI Arabic Stemmer to determine the roots of the Arabic words [18]-[19]. For the same example shown above, the ISRI stemming process is as follows :

"خط" "وصل" "جمع" "أداة" "عد" "حسب " " ربط " " سط"

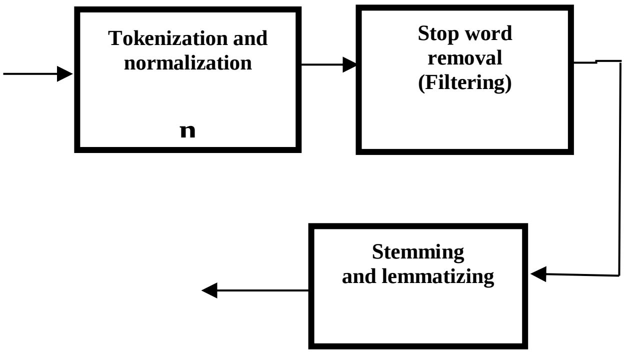


Figure 2. Pre-processing stages.

3.2 Arabic WordNet

As mentioned earlier, the WordNet is a lexical database which groups the words into sets of synonyms called synsets, along with the relations among these synonym sets, as finding a lexical resource offers broad coverage of the general lexicon of each word in the student’s answer which is extracted from the previous stage to define all the words that have a similar meaning. We have used the AWN which is a multi-lingual concept dictionary that maps between word senses in Arabic and those in the Princeton WordNet that was expanded in 2016 [7]. We used the AWN in this research to find all the words that are synonymous with the student's answer; to increase the likelihood of the student’s correct answer which was used after the pre-processing step. For the same example shown above , the result of using AWN is shown in Table 1.

Table 1. Example of WordNet.

word	synonyms
وصل	بلغ , علم , نهى ....
وسط	قصد, جزع, قطع ....
ربط	وثق , شد ....
جمع	حقن, قرن , لصق, ألف, حفظ, حشد,.................

3.3 F-score for Feature Selection

The support vector machine is represented by sparse vectors under the vector space, where each word in the vocabulary is mapped to one coordinate axis. It is used on data to train a linear classifier which is characterized by the normal to the hyperplane dividing positive and negative instances [20], [21], [22]. We apply feature selection aiming to pre-define the number of the highest scoring features to be included in a classifier by using the F-score technique. F-score is a feature selection method in SVM, which identifies the differences between two classes (positive and negative). The value of F-score for each feature is computed using Equation (1) [4]:

$$F(i) = \frac{\left(\overline{x_i^+} - \overline{x_i}\right)^2 + \left(\overline{x_i^-} - \overline{x_i}\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \overline{x_i^{(+)}}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \overline{x_i^{(-)}}\right)^2} \tag{1}$$

where *K* is a positive or negative instance, *xi* ¯, *xi +:* the average of i feature positive and negative dataset. *k, i*: the jth feature of the ith positive /negative instance. After determining the score of each feature, we then obtain the threshold value through calculating the average of F-score for all features. If the value of F-score is bigger than the mean value of all F-scores, the feature is added to the feature space, whereas if the value of F-score is less than the mean value of all F-scores, the feature is removed from the feature space. F-score is used to decide or select the feature that affects the score of student's answer which determines the positive and negative; according to a related or un-related answer. As shown in Figure 3, the related answers (positive) are taken, but the others (negative) are ignored.

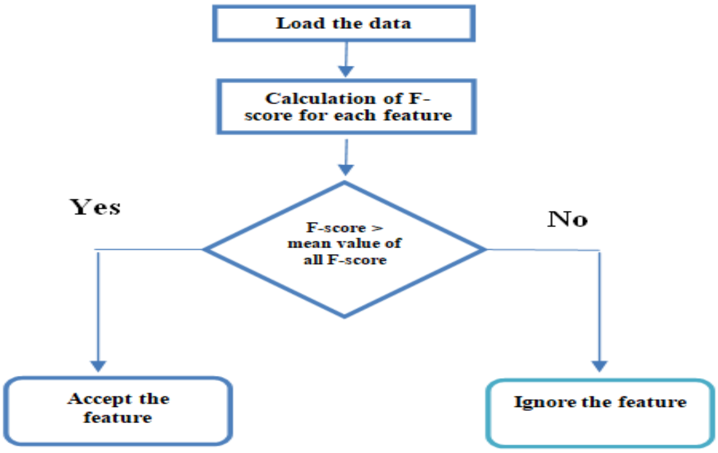

Load the data
Calculation of F-score for each feature
Yes
F-score > mean value of all F-score
No
Accept the feature
Ignore the feature

Figure 3. F-score feature selection.

3.4 Cosine Similarity

Cosine similarity is used to measure the cosine of the angle between any two vector spaces. It can be seen as "a comparison between documents" on a normalized space, as we are not taking into account the weight of each word count for each document, but the angle between documents. Cosine similarity will generate a value which articulates how correlated the two documents are by considering the angle as an alternative of the magnitude [7,22]. Cosine similarity is computed using Equation (2):

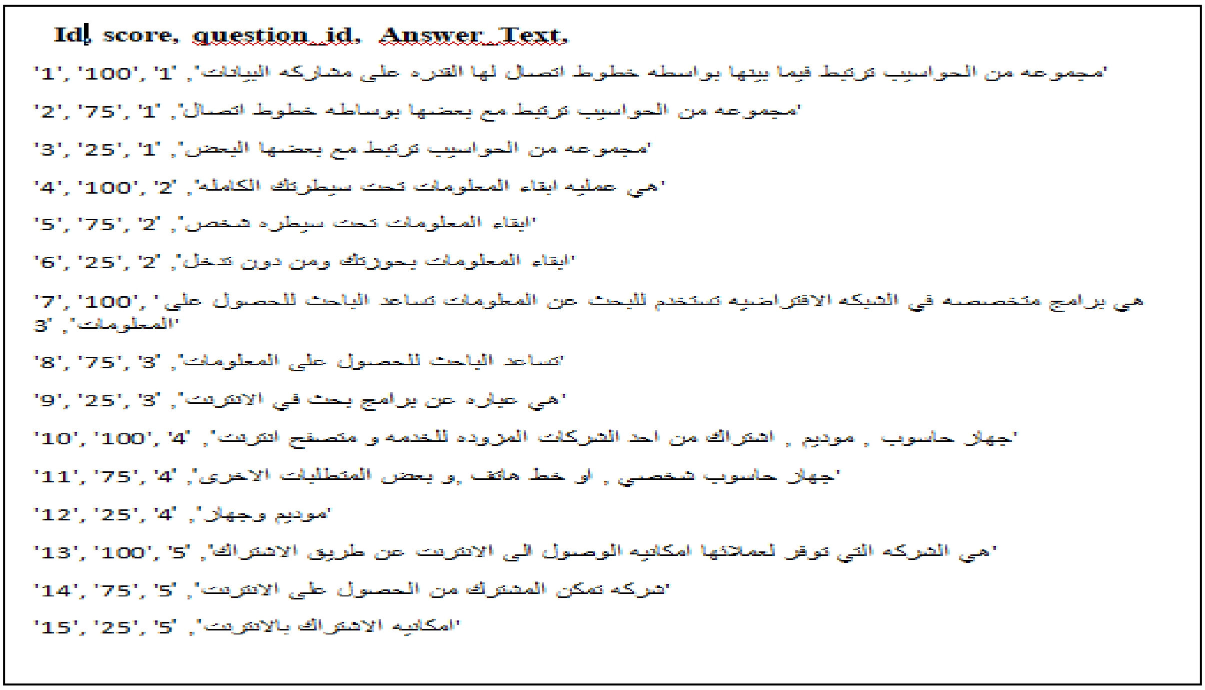
$$\cos\theta = \frac{\sum_i w_{q,i} . w_{i,j}}{\sqrt{\sum_j w_j^2} . \sqrt{\sum_j w_{i,j}^2}} \tag{2}$$

4. Experimental Design and Results

To evaluate the proposed model effectively, we carried out in this paper a comparative analysis of the impact of Arabic WordNet in automated essay grading. The dataset used is created in MYSQL as a CSV file, as data is collected from a computer, science and social school lectures from Allu'lu'a modern school in Madaba-Jordan, containing 120 questions (a sample of the questions used is shown in Table 2) along with 3 classes of model answers for each question and for each question with 30 student sample answers, a sample of model answers with human score is shown in Figure 4. The dataset was designed after the Hewlett Foundation Automated student assessment prize. The experiments were divided into two stages, where the first stage is the use of the proposed model without AWN; the cosine similarity degree for our proposed model and the human score are shown in Table 3.

Table 2. Sample of questions in the dataset.

Question ID	Question Text
1	عرف الإنترنت
2	ماذا يقصد بأمن المعلومات؟
3	ماذا يقصد بمزود خدمة الإنترنت؟
4	ماذا نعني بمحركات البحث؟
5	ما هي متطلبات الاتصال بالإنترنت؟
6	ماذا يقصد بمزود خدمة الإنترنت؟
7	ما هي وظيفة جهاز المودم؟
8	اذكر خدمات البريد الإلكتروني
9	علل احترام حقوق الملكية الفكرية عند استخدام الإنترنت؟
10	عرف بروتوكول الإنترنت
11	عرف الصراع الاجتماعي
12	عرف المنصهر
13	عرف البوصلة
14	عرف الكوكب
15	اذكر مكونات الدارة الكهربائية



Id, score, question_id, Answer_Text,

'1', '100', '1' ,'مجموعه من الحواسيب ترتبط فيما بينها بواسطه خطوط اتصال لها القدره على مشاركه البيانات'

'2', '75', '1' ,'مجموعه من الحواسيب ترتبط مع بعضها بوساطه خطوط اتصال'

'3', '25', '1' ,'مجموعه من الحواسيب ترتبط مع بعضها البعض'

'4', '100', '2' ,'هي عمليه ابقاء المعلومات تحت سيطرتك الكامله'

'5', '75', '2' ,'ابقاء المعلومات تحت سيطره شخص'

'6', '25', '2' ,'ابقاء المعلومات بحوزتك ومن دون تدخل'

'7', '100', 'هي برامج متخصصه في الشبكه الافتراضيه تستخدم للبحث عن المعلومات تساعد الباحث للحصول على
المعلومات', '3'

'8', '75', '3' ,'تساعد الباحث للحصول على المعلومات'

'9', '25', '3' ,'هي عباره عن برامج بحث في الانترنت'

'10', '100', '4' ,'جهاز حاسوب , موديم , اشتراك من احد الشركات المزوده للخدمه و متصفح انترنت'

'11', '75', '4' ,'جهاز حاسوب شخصي , او خط هاتف ,و بعض المتطلبات الاخرى'

'12', '25', '4' ,'موديم وجهاز'

'13', '100', '5' ,'هي الشركه التي توفر لعملائها امكانيه الوصول الى الانترنت عن طريق الاشتراك'

'14', '75', '5' ,'شركه تمكن المشترك من الحصول على الانترنت'

'15', '25', '5' ,'امكانيه الاشتراك بالانترنت'

Figure 4. Sample of model answers.

The second stage of the proposed model uses the AWN; the cosine similarity degree for our proposed model and the human score are shown in Table 4.

Table 3. Result of the proposed model without WordNet.

Question(id)	**human score**	**cosine result without WordNet**
1	1	0.94
2	1	0.94
4	0.75	0.67
5	0	0
11	0.75	0.66
12	0.75	0.82
15	0.25	0.21
30	0.25	0.22
36	0.25	0.1
40	1	0.97

Table 4. Score result of the proposed model using WordNet.

Question(id)	**human score**	**cosine result with WordNet**
1	1	0.98
2	1	0.98
4	0.75	0.8
5	0	0
11	0.75	0.67
12	0.75	0.85
15	0.25	0.24
30	0.25	0.3
36	0.25	0.21
40	1	0.98

The result of experiments demonstrates that the accuracy of the proposed model with AWN is close to human score as illustrated in Figure 5.

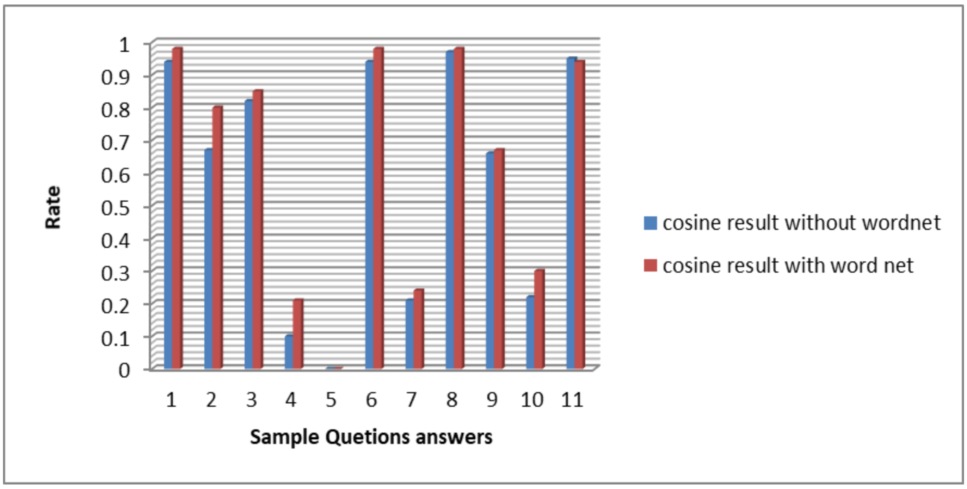


Figure 5. Rates of cosine similarity with and without AWN.

Moreover, to evaluate the effect of using AWN in the proposed model, a comparison was conducted between the human score and the score produced by the proposed model for a student's answer, using Mean Absolute Error (MAE) value and Pearson Correlation Coefficient. The MAE of the proposed model with the use of Arabic WordNet is 0.117 less than MAE of the proposed model without using Arabic WordNet. So, this result indicates that the proposed model will improve in the Arabic Automated Essay System with AWN, as shown in Figure 6.

Moreover, we used Pearson correlation which is a statistical measure that is used to determine whether or not there is a correlation between the scores produced by the proposed model and the human score.

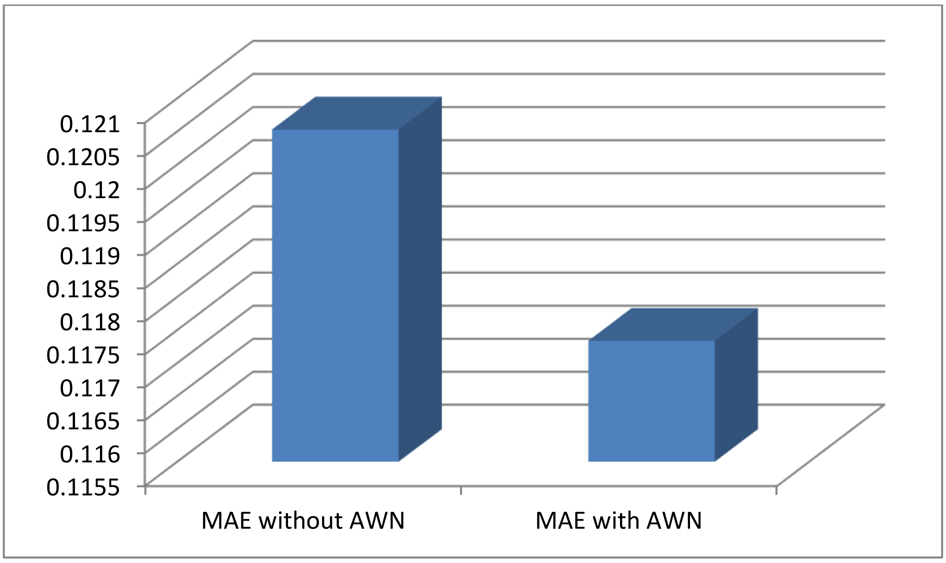


Figure 6. MAE.

Accordingly, the Pearson correlation result for the proposed model compared to human score is between 0.5 and 1, as it shows a high positive correlation that represents having the best correlation magnitude, as shown in Figure 7.

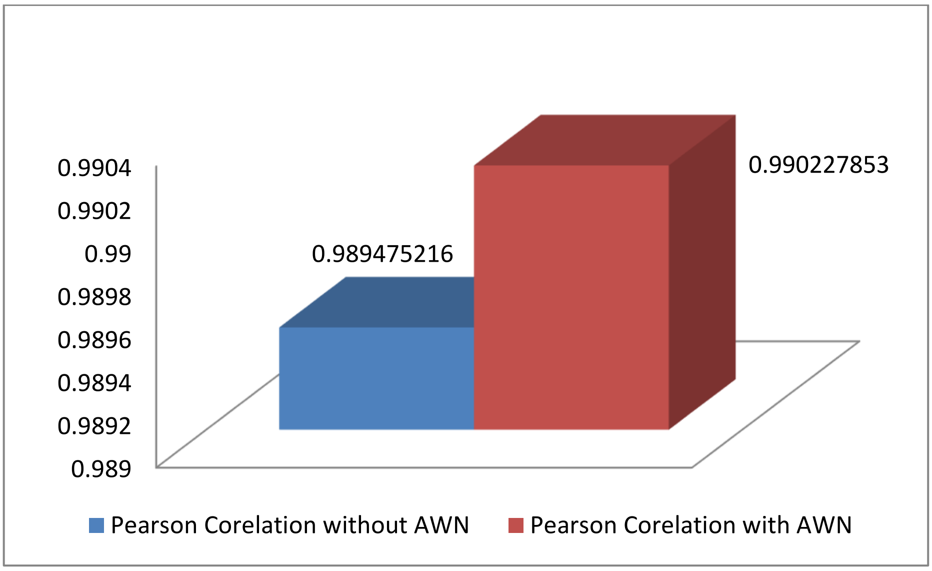


Figure 7. Pearson correlation.

5. Conclusion

This paper presented an Automated Arabic essay grading model to achieve better accuracy, while studying the role of Arabic WordeNet and F-score as efficient tools to extract features from the students' answers and from the model answers. We used cosine similarity to compute the score for the students. The focus of this work was to enhance the accuracy of the automated essay system to match human score by adding Arabic WordNet. The dataset created contains 120 questions with 3 model answers for each question. In addition, the dataset used in this paper was created according to Hewlett Foundation ASAP standards from Kaggle datasets. The automated essay grading results showed that the proposed model coupled with using Arabic WordNet (AWN) produces a better result compared to the case without using Arabic WordNet according to mean absolute error value and Pearson correlation. Based on the outcome of this research, there is an outlook for future work on using machine learning and neural network models to enhance the accuracy of Arabic essay grading along with studying the impact of the word-embedding technique.

REFERENCES

[1] D. Surya, V. Madala, A. Gangal, S. Krishna, G. Anjali and S. Ashish, "An Empirical Analysis of Machine Learning Models for Automated Essay Grading," PeerJ Preprints, pp. 1-14, India, 2018.

[2] V. Ramalingam, A. Pandian, P. Chetry and H. Nigam, "Automated Essay Grading Using Machine Learning Algorithm," Journal of Physics, Conference Series, vol. 1000, p. 012030, [Online], Available: 10.1088/1742-6596/1000/1/012030, 2018.

[3] M. Lilja, Automatic Essay Scoring of Swedish Essays Using Neural Networks, Statistics Uppsala University, 2018.

[4] A. Suresh and M. Jha, "Automated Essay Grading Using Natural Language Processing and Support Vector Machine," International Journal of Computing and Technology, vol. 5, no. 2, 2018.

[5] S. Gunes, K. Polat and S. Yosunkaya, "Multi-class F-Score Selection Approach to Classification of Obstructive Sleep Apnea Syndrome," Expert Syst. Appl., vol. 37, pp. 998-1004, 2010.

[6] W. H. Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," International Journal of Computer Applications, vol. 68, no. 13, 2013.

[7] Y. Regragui, L. Abouenour, F. Krieche, K. Bouzoubaa and P. Rosso, "Arabic WordNet: New Content and New Applications," Proceedings of the 8th Global WordNet Conference, pp. 330-338, 2016.

[8] A. Ewees, A. Mohammed Eisa and M. M. Refaat, "Comparison of Cosine Similarity and kNN Automated Essay Scoring," International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 13, pp. 8669-8673, 2014.

[9] M. Alghamdi, M. Alkanhal, M. Al-Badrashiny, A. Al-Qabbany, A. Areshey and A. Alharbi, "A Hybrid Automatic Scoring System for Arabic Essays," AI Communications, vol. 27, no. 2, pp. 103-111, 2014.

[10] R. Mezher and N. Omar, "A Hybrid Method of Syntactic Feature and Latent Semantic Analysis for Automatic Arabic Essay Scoring," Journal of Applied Sciences, vol. 16, no. 5, 2016.

[11] M. F. Al-Jouie and A. M. Azmi, "Automated Evaluation of School Children Essays in Arabic," Proc. of the 3rd International Conference on Arabic Computational Linguistics, vol. 117, pp.19-22, 2017.

[12] A. R. Abbas and A. S. Al-qaza, "Automated Arabic Essay Scoring (AAES) Using Vector Space Model (VSM)," Journal of Al-Turath University College, vol. 15, pp. 25-39, 2014.

[13] R. Martinez, H. H. Dong and D. Lee, "Automated Essay Scoring System by Using Support Vector Machine," International Journal of Advancements in Computing Technology (IJACT), vol. 5, no. 11, pp. 316-322, 2013.

[14] T. F. Gharib, M. B. Habib and Z. T. Fayed, "Arabic Text Classification Using Support Vector Machines," International Journal of Computers and Their Applications, vol. 16, no. 4, pp. 192-199, 2009.

[15] S. Alsaleem, "An Automated Arabic Text Categorization Using SVM and NB," International Arab Journal of e-Technology, vol. 2, no. 2, pp. 124-128, 2011.

[16] S. S.Aljameel, J. D. O'Shea, K. A. Crockett and A. Latham, "Survey of String Similarity Approaches and the Challenging Faced by the Arabic Language," Proceedings of 11th International Conference on Computer Engineering and Systems, Cairo, Egypt, 2016.

[17] A. Shehab, M. Faroun and M. Rashad, "An Automatic Arabic Essay Grading System Based on Text Similarity Algorithms," International Journal of Advanced Computer Science and Applications, vol. 9, no. 3, 2018.

[18] E. F. Al-Shalabi, "An Automated System for Essay Scoring of Online Exams in Arabic Based on Stemming Techniques and Levenshtein Edit Operations," arXiv preprint, 2016.

[19] B. Al-Shargabi, W. Al-Romimah and F. Olayah, "A Comparative Study for Arabic Text Classification Algorithms Based on Stop Words Elimination," Proceedings of the International Conference on

Intelligent Semantic Web-services and Applications (ACM), 2011.

[20] B. Al-Shargabi, F. Olayah and W. W. Romimah, "An Experimental Study for the Effect of Stop Words Elimination for Arabic Text Classification Algorithms," International Journal of Information Technology and Web Engineering (IJITWE), vol. 6, no. 2, pp. 68-75, 2011.

[21] G. Bilal and N. Rasha, "Semantic Analysis Based Customer Reviews Feature Extraction," Journal of Babylon University/Pure and Applied Sciences, vol. 25, no. 3, 2017.

[22] M. H. Nnguyen and F. Torre, "Optimal Feature Selection for Support Vector Machine," Pattern Recognition, vol. 43, pp. 584–591, 2014.

ملخص البحث:

في هذا البحث، يتم تصميم نظام مؤتمت لوضع الدرجات لاستخدامه في الجامعات والشركات والمدارس، يعتمد على الذكاء الاصطناعي وتقنيات المعالجة القائمة على اللغات الطبيعية. ويتميز النظام بقدرته على تحسين نظام وضع الدرجات من حيث التغلب على المشكلات المتعلقة بالتكلفة والوقت والجهد المبذول من المعلّم عندما يقوم بتصحيح إجابات الطلبة وأوراقهم. ويعود السبب في الانتشار الواسع لأنظمة وضع الدرجات المؤتمتة الى أمور تتعلق بالتكلفة والموثوقية والمعايير والتقنيات، وهي أمور تجعل النظام صالحاً للتطبيق والاستخدام لنصوص بلغاتٍ متعددة مثل الإنجليزية والفرنسية، ولغات أخرى. من جهة أخرى، لم يتم سوى إجراء عدد محدود من البحوث والدراسات في مجال أتمتة وضع الدرجات للنصوص العربية.

لذا، تقترح هذه الورقة نظام مؤتمت لوضع الدرجات لنصوص اللغة العربية. ويرتكز النظام المقترح على استخدام درجة ف (F-score) لاستخراج الخصائص من إجابات الطلبة ومن الإجابات النموذجية، جنباً الى جنب مع استخدام شبكة "ووردنت" العربية (AWN) كطريقة قيّمة مبنية على المعرفة من أجل التشابه في دلالات الألفاظ. ويتمثل الغرض من استخدام الشبكة العربية في إيجاد جميع الكلمات ذات العلاقة من إجابات الطلبة لإعطاء الطالب درجة على إجابته. وفي هذه الحالة، لن يتعرض الطلبة للإجحاف بدرجاتهم في الحالات التي لا يكتبون فيها الإجابة النموذجية الدقيقة؛ الأمر الذي من شأنه أن يقود الى تحسين نظام وضع الدرجات بحيث يتلاءم مع وضع الدرجات اليدوي الذي يقوم به الإنسان. وقد جرى تقييم النموذج المقترح في هذه الدراسة باستخدام مجموعة بيانات لنصوص باللغة العربية. وبينت النتائج أن النموذج المقترح يعطي نتائج تتوافق مع وضع الدرجات الذي يقوم به الإنسان.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 05, No. 03, December 2019.

DEEP LEARNING-BASED RACING BIB NUMBER DETECTION AND RECOGNITION

Yan Chiew Wong[1], Li Jia Choi[1], Ranjit Singh Sarban Singh[1], Haoyu Zhang[2] and A. R. Syafeeza[1]

(Received: 10-Jul.-2019, Revised: 27-Aug.-2019, Accepted: 28-Aug.-2019)

ABSTRACT

Healthy lifestyle trends are getting more prominent around the world. There are numerous numbers of marathon running race events that have been held and inspired interest among peoples of different ages, genders and countries. Such diversified truths increase more difficulties to comprehending large numbers of marathon images, since such process is often done manually. Therefore, a new approach for racing bib number (RBN) localization and recognition for marathon running races using deep learning is proposed in this paper. Previously, all RBN application systems have been developed by using image processing techniques only, which limits the performance achieved. There are two phases in the proposed system that are phase 1: RBN detection and phase 2: RBN recognition. In phase 1, You Only Look Once version 3 (YOLOv3) which consists of a single convolutional network is used to predict the runner and RBN by multiple bounding boxes and class probabilities of those boxes.YOLOv3 is a new classifier network that outperforms other state-of-art networks. In phase 2, Convolutional Recurrent Neural Network (CRNN) is used to generate a label sequence for each input image and then select the label sequence that has the highest probability. CRNN can be straight trained from sequence labels such as words without any annotation of characters. Therefore, CRNN recognizes the contents of RBN detected. The experimental results based on mean average precision (mAP) and edit distance have been analyzed. The developed system is suitable for marathon or distance running race events and automates the localization and recognition of racers, thereby increasing efficiency in event control and monitoring as well as post-processing the event data.

KEYWORDS

Racing Bib Number, You Only Look Once Version 3, Convolutional Recurrent Neural Network.

1. INTRODUCTION

Over the past decades, great popularization of intelligence devices and rapid development of technology have brought forth enormous new outcomes and services that have prompted the huge demand of practical computer vision technologies. As opposed to the scanning of documents, natural scene text localization [1] and recognition contribute a method to directly access and utilize the textual data in the wild, which is apparently the most pressing technology. Nowadays, healthy lifestyle trends are more prominent around the world. Such new trends help organize running activities in order to inspire the awareness of the public of the importance of health. Large numbers of marathon running races in different formats for different situations have been held and inspired interest among people of different ages, genders and countries. Such diversified truths increase more difficulties to comprehending marathon video or images, since such process is often done manually, which yields a process made difficult by the sheer number of available photos. There are multiple methods for text localization and recognition in natural scene images and videos; for instance: parking, border control and analysis of traffic flow. However, these approaches are not good enough to obtain accurate and precise results for RBN localization in marathon images, because they usually rely on characteristics of rich texts, but not numerals [2]. The runners or participants have a single racing bib number (RBN) to indicate themselves and this RBN is usually presented on heterogeneous backgrounds and different materials. Such tag number is generally printed on a cardboard tag and pinned onto T-shirts of different colours and at different parts of each competitor’s body during the marathon race. There are some problems faced in the application of RBN detection and recognition. Firstly, competitors run

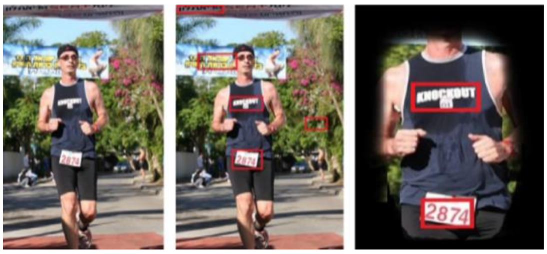
1. Y. C. Wong*, L. J. Choi, S. S. S. Ranjit and A. R. Syafeeza are with Centre for Telecommunication Research & Innovation (CeTRI), Fakulti Kejuruteraan Elektronik dan Kejuruteraan Komputer (FKEKK), Universiti Teknikal Malaysia Melaka (UTeM), Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia. Email: ycwong@utem.edu.my
2. H. Zhang is with Zhejiang Ocean University, School of Naval Architecture and Mechanical -Electrical Engineering, Zhoushan City, Zhejiang, China. Email: haoyu19871202@163.com

together with different speeds and background complexity varies continuously with moving objects, sky, buildings, trees, among others. Secondly, input images captured in natural scene are usually influenced by occlusion and the loss of information or quality. Thirdly, inconsistent light intensities can influence the images captured, which results in unbalanced illumination throughout the images. Hence, achieving accurate performance for localizing and recognizing RBN in marathon races is a challenging and difficult task.

Along with that, the security of running events is the priority concern for organizers of distance running events. A tragedy incident, Boston Marathon bombing, happened on 15 April 2013. Apart from the extensive video footage and photos from local surveillance systems and witnesses' smartphones and cameras, local, state and federal law enforcement organizations failed to indicate the suspects of this incident by using only face detection technology. As a result, the proposed system in this project can assists in studying and comprehending such images to extract the desired information regardless of age, gender and country. Furthermore, the proposed system of RBN detection and recognition has the capability to assist gait analysis. Gait analysis is the study of motion figures of lower limbs, such as weight of individual, footwear, length of limbs and postures integrated with motion. It can also be employed as a biometric calculation to recognize known individuals and discover unknown subjects. Gait analysis is widely implemented to recognize the performances of athletes in sport training or events such as running and swimming, in order to ensure that the athletes have worn the correct footwear and are performing foot movement that won't result in injury to them. Runner detection included in this proposed system will certainly provide the desired information needed for gait analysis despite of RBN recognition only.

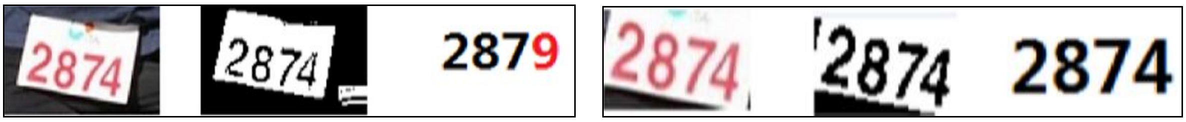
There are many online resources available selling photos of individuals, since runners may want to purchase their photos captured during marathon events as personal memories. Generally, runners shall find and purchase their respective photos by using their RBN. However, there are thousands of images captured by organizers and photographers attending marathon running races, which results in high time consumption and patience needed in sorting such images according to the RBN of each runner. As a result, the proposed system is able to overcome this problem by localizing and recognizing RBN automatically, hence increasing the efficiency in sorting marathon images.

There are a few previous research studied on RBN recognition using image processing methods. Figure 1 is the method proposed by P. Shivakumara using SVM and multiple image processing methods based on 200 plus images [3]. It illustrates the inability of scene text detection methods for RBN detection. Figure 1(i) shows an input image captured from a running race. Figures 1(ii) and (iii) show RBN localization with and without detecting torso. A lot of false positives are detected by text detection method without torso detection, while the desired output is achieved by including torso detection before text detection. Figure 2 shows the result of binarization and the result of RBN recognition with and without torso detection. It can be observed that incorrect binarization and recognition results are obtained without torso detection. Therefore, the accuracy and precision of RBN rely significantly on the text detection methods. P. Shivakumara method [3] combines torso and text detection methods. It is not related to runner movement directions and does not require face information as the method proposed by Ami et al. [4] who used the stroke width transform (SWT) method [5]-[6] to extract characters in input images, where the characters with similar stroke width are grouped together for producing text region. Then, face detection method is applied in order to detect the face of the runner. At the end, the detected torso is used to indicate and recognize the true RBN by using Tesseract OCR engine [7]. The limitation of such method is tremendously large numbers of runners running with different speeds during running races such as marathon races and hence it is very difficult to localize and detect each runner's face. N. Boonsim [8] applied edge-based technique [9]-[10] to extract edges of an input image captured during a marathon race and morphological operations in mathematical form were used to combine the edges in order to generate RBN area and fade other areas. Runner detection method is used to detect the face of the runner first and is then extended to the torso of runner. The detected runner is used to indicate the position of RBN. The verification process applies image intersection between the candidate region image and the runner's body image. After that, image contrast enhancement is employed to enhance the contrast of the image. Local contrast improvement method [9] is employed to refine local contrast instead of global contrast due to that it has the capability to solved images with complicated backgrounds. Text localization and recognition methods alone have failed to obtain high performance for RBN detection and recognition because of background variations and uncontrolled issues. Inspired by these observations, a system which uses the method of runner detection to reduce background complexity and perform RBN detection and recognition based on deep learning is proposed in this work.



(i) (ii) (iii)

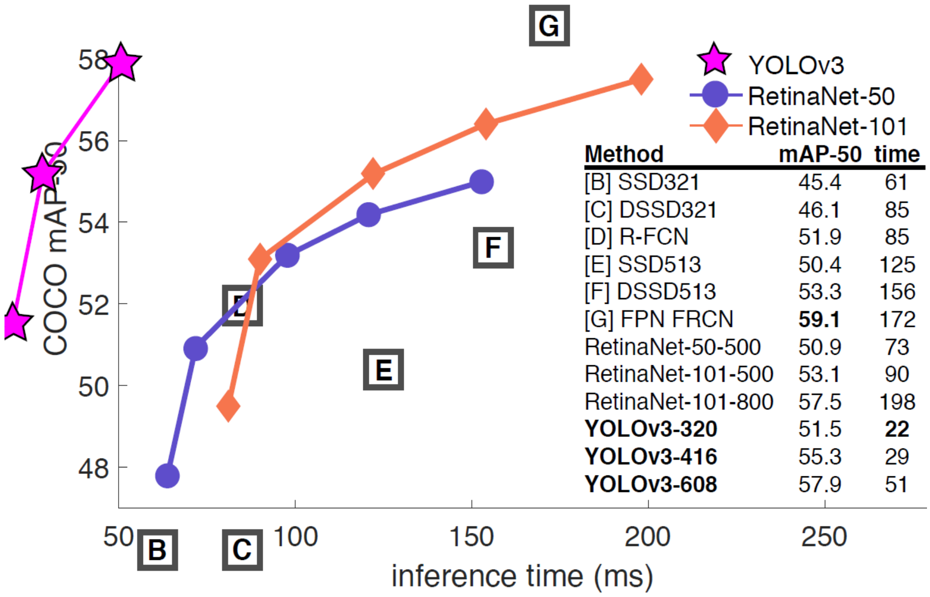
Figure 1. Scene text detection method for RBN detection [3] (i) images of running races (ii) inability of scene text detection method for RBN detection without detecting torso and (iii) RBN localization after torso detection.



(i) (ii)

Figure 2. Binarization and recognition results for input images (i) without torso detection (ii) with torso detection [3].

In this work, an artificial intelligence cascade network, which can automatically detect and recognize RBN during marathon races, will be implemented on a graphic processing unit (GPU) by using deep learning. This is the first RBN recognition system implemented by using deep learning. Cascaded network topology, which consists of an object detection algorithm: You Only Look Once v3 (YOLOv3) [12]-[13] and an object-based sequence recognition algorithm: Convolutional Recurrent Neural Network (CRNN), will be developed. YOLOv3 is a pre-processing step that transforms an input image quickly into a sequence of image features for sequence-like object detection through CRNN. YOLOv3 operates dramatically faster than other recent detection methods. As shown in Figure 3, at 320 x 320, YOLOv3 runs in 22ms at 28.2mAP as accurate as Single Shot Detection (SSD), but three times faster.



Method	mAP-50	time
[B] SSD321	45.4	61
[C] DSSD321	46.1	85
[D] R-FCN	51.9	85
[E] SSD513	50.4	125
[F] DSSD513	53.3	156
[G] FPN FRCN	**59.1**	172
RetinaNet-50-500	50.9	73
RetinaNet-101-500	53.1	90
RetinaNet-101-800	57.5	198
YOLOv3-320	51.5	**22**
YOLOv3-416	55.3	29
YOLOv3-608	57.9	51

Figure 3. YOLOv3 runs significantly faster than other detection methods with comparable performance [13].

The critical characteristic of bib recognition is that bib sizes may change randomly. As a result, the most famous deep models, such as Deep Convolutional Neural Network (DCNN) [16], are not suitable to apply directly to object-based sequence recognition after the target text has been detected by using a text detector [17]. This is due to that DCNN models generally operate on inputs and outputs with certain dimensions and hence such methods are unable to generate an alterable-length label sequence. Recurrent Neural Network (RNN) is particularly used for recognizing sequences. The main benefit of RNN is that it does not require the location of each component in a sequence object image in both phases of training and testing. CRNN is a combination network of DCNN and RNN. For sequence-like objects, CRNN can be directly trained from sequence labels such as words without any annotation of characters and learning informative data straightly from input images which do not need any hand-craft features or pre-processing steps such as binarization, segmentation and character localization as needed in conventional image processing methods proposed by previous research work.

YOLOv3 will be used for detecting the runner and bib number by multiple bounding boxes. CRNN will be used for recognizing the bib number through generating a label sequence for each input image and then selecting the label sequence that has the highest probability. The work follows with analyzing and enhancing the performance of the system in order to achieve high accuracy and precision. The text language is limited to the representative tag number of the runner only, where the font size and type of the text dataset typically depend on the resources of such dataset. Name of logo and name of runner are not considered in this work.

2. Methodology

2.1 Phase 1- YOLO v3

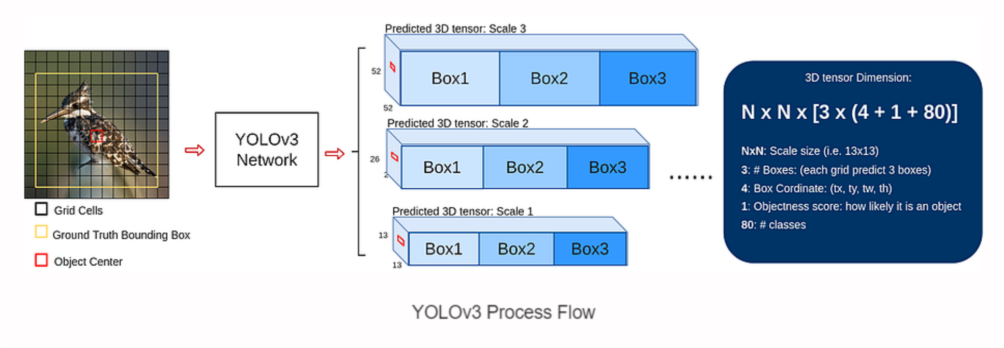
YOLO [12] uses single regression to detect the target object directly from image pixels by predicting multiple bounding boxes and class probabilities for those boxes. YOLO trains on full images and straight away optimizes detection performance. YOLOv3 [13] is the third object localization algorithm in the family. There are no fully-connected layers or pooling layers in the network architecture of YOLOv3. Therefore, YOLOv3 has the capability to handle images with variable size.

In this work, the first dataset used to train YOLOv3 has the total amount of 4096 images consisting of 1070 pixels as width and 1600 pixels as height in PNG format and is created by collecting images one by one from an online database [11]. Such images are captured at places near to the finish line during Delaware marathon running race in 2018. Then, a graphical image annotation tool is used in a process named as labelimg. Labelimg tool is a graphical image annotation tool. It is written in Python language and utilizes qt library for its graphical interface. After installing its dependencies, the tool is run to perform the image annotation for the 4096 images. Every image is provided with ground truth values which indicate the detail annotation and temporal localization of each of the target objects that are: runner, bib and number. Therefore, a text file will be generated after saving each of the images in YOLO format. Each row in the text file represents a single bounding box in an input image. The format of the bounding box is shown in Table 1.

Table 1. Format of the bounding box in YOLO format.

<object-class-id><center-x><center-y> <width> <height>	**Object-class-id** is a parameter which represents the class index of an object. It ranges from zero to (number of classes – 1).
x: x-coordinate (in pixels) of the center of the bounding box	**<center-x>** and **<center y>** are the respective coordinates of the center of the bounding box. Each of the values have been normalized by the width and height of the input image.
y: y-coordinate (in pixels) of the center of the bounding box	
w: width (in pixels) of the bounding box	**<width>** and **<height>** represent the normalized width and height of bounding boxes as follows: <center-x> = $\frac{x}{W}$, <center-y> = $\frac{y}{H}$,<width> = $\frac{w}{W}$, <height> = $\frac{h}{H}$
h: height (in pixels) of the bounding box	
W: width (in pixels) of the whole image	

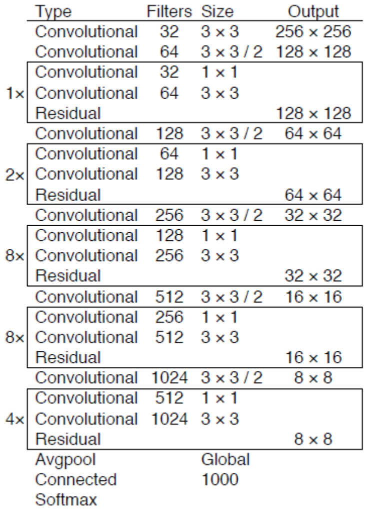
During training, YOLOv3 is applied with input images to predict 3D tensors that act as the last feature map corresponding to three scales that are regions 82, 94 and 106, as shown in Figure 4. Such three scales are constructed to localize target objects with different sizes. For instance, a scale of 25 x 25 means that an input image will be split into 25 x 25 grid cells. Every single grid cell relates to a 1 x 1 x 255 voxel inside a 3D tensor. 255 is calculated from formula of 3 x (4+1+80). By referring to Figure 4(iii), the values in a 3D tensor are shown, which are composed of confidence score, class confidence and bounding box coordinates.


Grid Cells
Ground Truth Bounding Box
Object Center
YOLOv3 Network
Predicted 3D tensor: Scale 3
52
52
Box1
Box2
Box3
Predicted 3D tensor: Scale 2
26
Box1
Box2
Box3
Predicted 3D tensor: Scale 1
13
13
Box1
Box2
Box3
......
3D tensor Dimension:
N x N x [3 x (4 + 1 + 80)]
NxN: Scale size (i.e. 13x13)
3: # Boxes: (each grid predict 3 boxes)
4: Box Cordinate: (tx, ty, tw, th)
1: Objectness score: how likely it is an object
80: # classes
YOLOv3 Process Flow

(i) (ii) (iii)

Figure 4. Process flow of YOLOv3 [12].

In phase 1, runner, racing bib and number are detected by using YOLOv3. The overall process of fetching the labeled images to YOLOv3 model is associated with using OpenCV library [14]. By applying the OpenCV library, all the input images are scaled to the same size. After resizing the input images, such images will be fed to YOLOv3. The network architecture of YOLOv3 is separated into several modules, as shown in Figure 5. First layer is the total of 32 convolutional layers, where the input images have been supplied. These convolutional layers are implemented to extract features from input images. Second layer consists of the residual layers. Such layers are proposed to vary the training process of the deep neural network from layer-by-layer training into phase-by-stage training. Therefore, the deep neural network is separated into few segments in order to realize the problem of gradient explosion or gradient dispersion of the network. Third layer is Darknet-53, which ranges from the 0th layer until the 74th layer. There are 53 convolutional layers and the remaining are the residual layers. Darknet-53 acts as a key component of the network architecture to extract features and implements a series of 3 x 3 and 1 x 1 convolutional layers. Last layer is the feature interaction layer of the YOLO network. It is separated into three scales that are regions 82, 94 and 106 in the feature pyramid network. Local feature interaction is implemented by convolutional kernel layers known as fully-connected layers in each region. It is used to obtain local feature interaction among feature maps and hence accomplish regression and classification.



	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Figure 5. Network structure of YOLO v3 [13].

2.2 Phase 2- CRNN

RBN text is recognized by using CRNN model. CRNN [15] is the combination network of DCNN [16] and RNN [18]. For sequence-like objects, CRNN is composed of several unique advantages, such as straight learning from sequence labels, image data for binarization, segmentation or component localization without the need of hand-craft features or pre-processing steps. It also possesses the same characteristic of RNN that makes it able to generate a sequence of labels. In spite of that, CRNN is unlimited to the sizes of sequence-like objects, but is only limited to the demand of height normalization in both phases of training and testing. Besides, it obtains excellent results on scene text recognition compared to other prior arts [17]. Lastly, it has low memory assumption by consisting of lesser parameters compared with a standard DCNN model.

Second dataset used to train for CRNN is composed of 100,000 images that contain digit numbers only in grey scale, as shown in Figure 6. These images are generated by a text generator python script. Types of font that are similar to the images of the first dataset are searched and downloaded from an online database and the path of font types is specified in the script. Types of font used to generate the second dataset include Identikal Sans Bold, Contax Bold, Roadway and Sofia Pro Semibold, and so on. Only one type of font is utilized to generate texts at one time while keeping switching to another font type to generate the dataset until it accumulates to a total of 100,000 images. In the text generator script, only digit numbers are specified and thus images composed of only digit numbers will be generated in random arrangement. Gaussian blur and rotate functions are applied in the script and hence some of the images are in the condition of blurriness or rotate slightly to assimilate practical real-life conditions, such as insufficient light source and indirectly facing the camera.

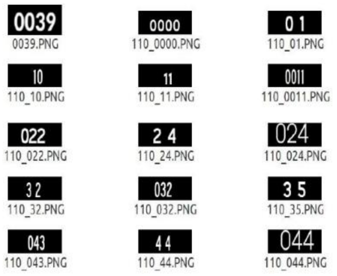


Figure 6. Second dataset generated.

According to Figure 7, the network architecture of CRNN is composed of three components which are from bottom to top convolutional layers, recurrent layers and a transcription layer. At the most bottom of CRNN, the convolutional layers automatically extract a sequential feature from each input image, while on top of the convolutional network, a recurrent network is constructed for making prediction for each frame of the feature sequence which is outputted from the convolutional layers. The transcription layer at the top of CRNN is utilized to translate the per-frame predictions by the recurrent layers into a label sequence.

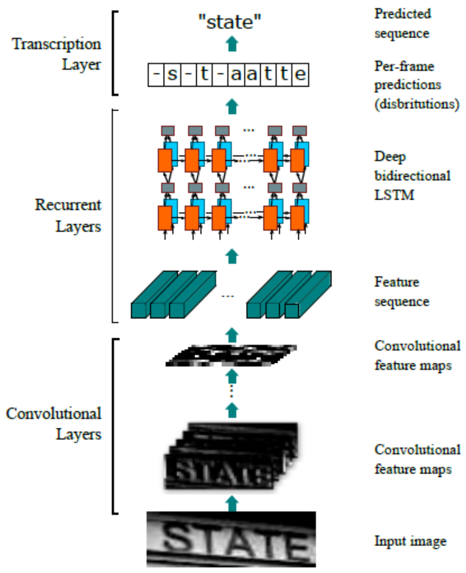


Figure 7. Network structure of CRNN [14].

Before starting training for the CRNN model, it is essential to convert the second dataset composed of 100,000 images into Lightning Memory-Mapped Database (LMDB) format. LMDB is a software library that supplies an embedded transactional database in key-value store format. Key-value store is considered as data storage built for saving, extracting and manipulating associative arrays. LMDB employs memory-mapped files and hence enhances the performance of input and output. A memory-mapped file is a division of virtual memory that has been appointed an undeviating byte-for-byte correlation with few sections of file or resource. As a result, LMDB is operating well when dealing with very large datasets. After separating the second dataset into training subset and testing subset, both subsets are converted into LMDB format. This will generate two files for respective training and testing subsets that are data.mdb and lock.mdb with different sizes.

There are few parameters needed to be adjusted before starting to train. The paths of training and testing images in LMDB format are specified and alphabet trained is set to ‘0123456789’ consisting of 10 classes. CUDA acts as a parallel computing platform and programming model that utilizes GPU for general purpose computing is enabled during training of CRNN model. Besides, several parameters utilized in the training algorithm are unchanged. These are batch size, number of GPU, learning rate, number of data loading workers, width and height of input images. Such parameters are unvaried during the training process while the number of epoch is changeable until the recognition output achieves correct results.

2.3 Cascading Both Models

After finishing collecting, annotating and generating the two datasets, all the images will then be synchronized and inserted into the proposed networks separately to be used as training and testing subsets. Training subset is utilized for learning so as to satisfy the parameters of the classifier. Input images have their respective ground truth files for supervised learning. Testing subset is utilized to evaluate the performance of a fully-trained classifier. No same image should be used as part for both of training and testing subsets. Both datasets have been split in the ratio of 7:3 as training subset to testing subset as shown in Table 2.

Table 2. Training and testing subsets.

Dataset	Training Subset	Testing Subset	Total
1 (RBN)	2868	1228	4,096
2 (Numbers)	700,000	300,000	100,000

The two pre-trained models trained from the two phases that are YOLOv3 for runners and RBN detection and CRNN for RBN recognition are combined together form a cascade network aimed for RBN detection and recognition based on deep learning. Model files, such as classes, model configuration and weight files for YOLOv3 and weight file of CRNN, are specified in a python script and will be parsed by using command line arguments. At the beginning in the script, the demanded packages that include OpenCV, NumPy, Pytorch and utils are imported. Parameters of confidence threshold, non-maximum suppression threshold, width and height of input image are adjusted. Confidence threshold and non-maximum suppression threshold are set to 0.5 and 0.4, respectively. Predicted bounding boxes consist of values lower than that will be discarded. When an input image is parsed to the cascade network, YOLOv3 model will be loading and obtaining the name of classes during training and drawing the predicted bounding box in rectangular form. Hence, confidence scores of each of the predicted bounding boxes are obtained and displayed at the tops of the boxes. Moreover, the predicted bounding boxes with classes of ‘number’ will be cropped by using OpenCV and parsed to the CRNN model. CRNN will be using CUDA to perform the RBN recognition process.

According to Figure 8, the architecture pipeline or network is consists of two phases: runner, racing bib and number detection in phase 1 and number recognition in phase 2. Therefore, two different networks are utilized which are YOLOv3 and CRNN to localize and recognize RBN. During phase 1, YOLOv3 consists of a single convolutional network used to predict RBN by multiple bounding boxes and class probabilities of boxes since there may be more than one RBN in one image. Since the target object detected is only the RBN, runner and racing bib detection is proceeded to avoid that YOLOv3 detects the regions out of the runner, such as background variation when the runners are running. Therefore, the detected RBN can be parsed into CRNN to undergo the process of recognition. For the second phase, CRNN is used to output a label sequence for each input image and then select the label sequence that has the highest probabilities. As a result, CRNN will output the contents of RBN detected.

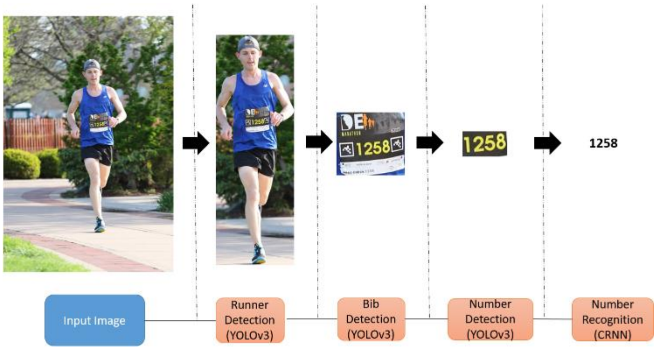

1258
Input Image
Runner Detection (YOLOv3)
Bib Detection (YOLOv3)
Number Detection (YOLOv3)
Number Recognition (CRNN)

Figure 8. Method selection.

2.4 Prediction

After the pre-trained models for the two phases are combined together. The validation and testing subsets will be utilized to predict the RBN localization and recognition. The predicted process is evaluated with combined pre-trained models to enhance the performance of the RBN detection and recognition system. By applying the testing subset, the cascade system has the capability of comparing the predicted bounding box with the actual bounding box for each input image.

2.5 Post-processing

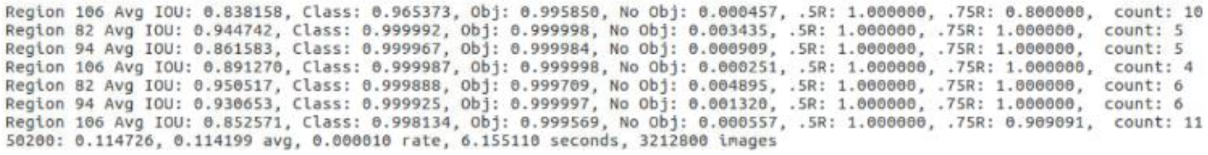
Post-processing is essential for YOLOv3 model only in order to localize RBN from testing images. The predicted bounding boxes from previous step are post-processed by using the method of non-maximum suppression which had been built in the deep neural network implementation of OpenCV. Non-maximum suppression is an important step of post-processing in many computer vision applications, especially for object detection. It is utilized to convert a smooth response map that activates many inaccurate object window hypotheses, which means only a single predicted bounding box for each target object.

3. Results and Discussion

There are two image datasets and each of them will be employed as inputs to train YOLOv3 and CRNN separately. Hence, the actions of localization and recognition of RBN will be acquired from images captured from marathon running events. All the precision results obtained from the models are enumerated to determine the exact location of RBN. Training loss, accuracy and precision of each network will be presented and analyzed in the following sub-sections.

3.1 Training YOLOv3 and CRNN with Own Datasets

In this project, there are two different datasets that are used for RBN detection and recognition during marathon events. In the training process, the parameters utilized in the training algorithm for both neural networks are very important in order to regulate the precision of localization and recognition output. Both neural networks are trained separately with their respective datasets. For YOLOv3, there are several parameters generated during training, as shown in Figure 9. The training process will keep updating the weights of YOLOv3 model from each iteration. Regions 82, 94 and 106 represent distinct sizes of parameters predicted at three distinct scales respectively in the feature pyramid network of YOLOv3. Convolutional layer of region 82 is the greatest prediction scale that utilizes a huge mask that makes it able to detect smaller object. Convolutional layer of region 94 is the medium prediction scale that utilizes a medium mask while convolutional layer of region 106 is the smallest prediction scale for localizing larger objects. Since the batch and subdivision are set as 64 and 16 respectively in the model configuration file, the training iteration consists of 4 groups of regions 82, 94 and 106, where each group consists of 16 images during training output. As a result, the model will simply choose 64 batches of images from all training sets to be fed into the model during each iteration. All of the image batches chosen will be separated into subdivision of 4 times in order to minimize memory consumption.



```
Region 106 Avg IOU: 0.838158, Class: 0.965373, Obj: 0.995850, No Obj: 0.000457, .5R: 1.000000, .75R: 0.800000,  count: 10
Region 82 Avg IOU: 0.944742, Class: 0.999992, Obj: 0.999998, No Obj: 0.003435, .5R: 1.000000, .75R: 1.000000,  count: 5
Region 94 Avg IOU: 0.861583, Class: 0.999967, Obj: 0.999984, No Obj: 0.000909, .5R: 1.000000, .75R: 1.000000,  count: 5
Region 106 Avg IOU: 0.891270, Class: 0.999987, Obj: 0.999998, No Obj: 0.000251, .5R: 1.000000, .75R: 1.000000,  count: 4
Region 82 Avg IOU: 0.950517, Class: 0.999888, Obj: 0.999709, No Obj: 0.004895, .5R: 1.000000, .75R: 1.000000,  count: 6
Region 94 Avg IOU: 0.930653, Class: 0.999925, Obj: 0.999997, No Obj: 0.001320, .5R: 1.000000, .75R: 1.000000,  count: 6
Region 106 Avg IOU: 0.852571, Class: 0.998134, Obj: 0.999569, No Obj: 0.000557, .5R: 1.000000, .75R: 0.909091,  count: 11
50200: 0.114726, 0.114199 avg, 0.000010 rate, 6.155110 seconds, 3212800 images
```

Figure 9. Output parameters produced during training of YOLOv3.

Parameter of ‘Avg IOU’ for the respective region represents the average IOU of the image in the present subdivision. It indicates the overlap of the predicted bounding box to the ground truth of target object and the union. The closer the value to 1, the better the bounding box prediction. For instance, ‘Avg IOU: 0.838158’ produced during training output has achieved quite high accuracy. ‘Class’ indicates the correctness of object classification to respective classes, where the values are expected to approach 1. ‘Obj’ is the average of the objectness or confidence score, which also calculates the probability of that there is an object in the bounding box, while ‘No Obj’ represents the opposite case. To produce a better result of localization, the value of ‘Obj’ shall be closer to 1 and ‘No Obj’ shall be approaching but not zero. ‘.5R’ and ‘.75R’ are referred to as the ratio of true positive objects predicted by the present model to the ground truth of object in each subdivision of images. If all objects are precisely predicted, then ‘.5R’ shall be equal to 1 and ‘.75R’ shall be equal to 0. ‘.5R’ is defined as the ratio with IOU greater than 0.5, while ‘.75R’ is the ratio with IOU greater than 0.75. ‘Count’ indicates the number of training images composed of true positive objects in all present subdivision images which is 4 in our case. By referring to Figure 9, since there are 4 or 6 true positive objects, this means that such division is composed of the images that do not belong in the localized object classes. Last row indicates the batch output, as shown in Figure 7. During the training process, every 1000 epochs will generate one weight file. As the number of epochs is increasing, the frequency of updating the values of weights will also increase until a minimum training loss is obtained.

CRNN model requires Connectionist Temporal Classification (CTC) [19] to do the training. CTC is specifically functional for neural networks composed of convolutional layers (CNN) to withdraw a chain of features and recurrent layers (RNN) to distribute information from such chains [20]-[21]. It generates character-scores for each chain component considered as a matrix in an easier way. By referring to Figure 10, there are two operations that need such matrix to proceed with calculating the loss value to train the neural network during training and decode the matrix to obtain the text contents from input images during inference. CTC benefits users by applying CTC loss function to learn the text contents existing in an input image and discard both location and width of characters in the input image. Furthermore, no further post-processing of the recognized text is required.

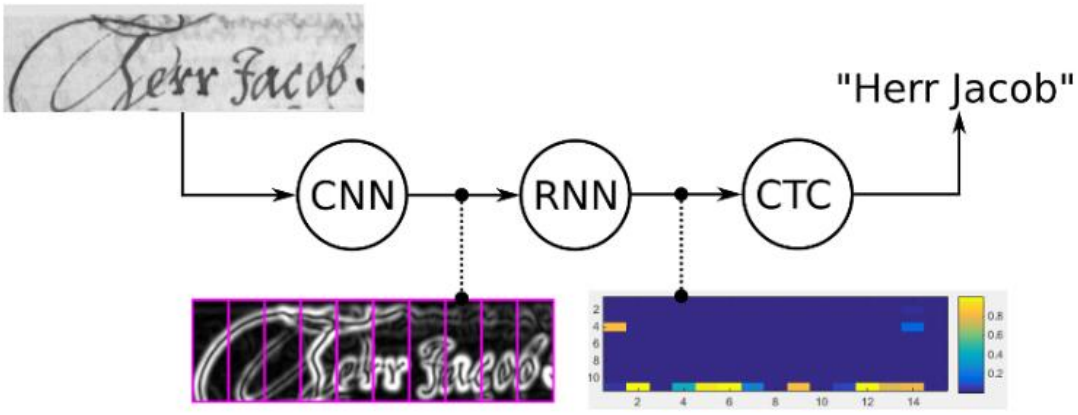
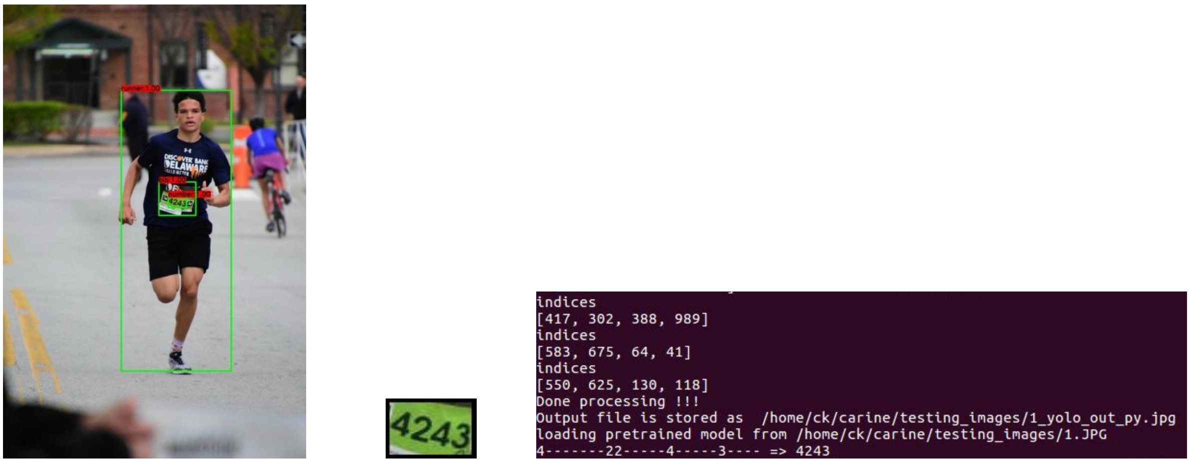


Figure 10. Process flow of neural network for handwriting recognition [15].

3.2 Result of Cascade Network

Since the targeted object is only limited to the RBN, the function of image cropping by using OpenCV is added in the script. Scripts utilized to run the models of YOLOv3 and CRNN are combined to become a cascade network. As a result, YOLOv3 will be used to detect runner, racing bib and number first, then the bounding box of the detected number will be cropped by using the crop function in OpenCV. The cropped image will be parsed to CRNN to undergo the recognition process as illustrated in Figure 11.



```
indices
[417, 302, 388, 989]
indices
[583, 675, 64, 41]
indices
[550, 625, 130, 118]
Done processing !!!
Output file is stored as  /home/ck/carine/testing_images/1_yolo_out_py.jpg
loading pretrained model from /home/ck/carine/testing_images/1.JPG
4-------22-----4-----3---- => 4243
```

(a) (b) (c)

Figure 11. Demonstration of RBN detection and recognition cascade network: (a) Objects detected using YOLOv3 model, (b) Bounding box of RBN cropped image (c) Recognition process using CRNN model.

3.3 Analyzing Predicted Result from Cascade Network

In this work, the prediction accuracy scores are acquired and evaluated for each of the classes localized by YOLOv3 model in the first phase. The metric utilized to implement the results of the temporal localization is named as mean Average Precision (mAP) for object classes, as shown in Table 3. Mean Average Precision (mAP) is usually employed to measure the accuracy of object detectors in terms of how well the predicted bounding box trained by the object detection model overlaps with the ground truth bounding box. The mAP metric is the product of recall and precision of the detected bounding boxes and its value ranges from 0 to 1. Precision determines how accurate the prediction is, while recall determines how well we find all the positives, as shown in Equations (2) and (3).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

Table 3. Mean average precision before and after fine-tuning with extra training images.

Dataset	Training Subset	Testing Subset	Total	Average Precision	
Training from scratch	2868	1228	4096	Runner	0.8958
				Bib	0.7164
				Number	0.6224
				mAP	0.7449
Fine-tuning	700	300	1000	Runner	0.9786
				Bib	0.8727
				Number	0.7032
				mAP	0.8515

The higher the value of mAP, the better the result. The mAP can be evaluated by calculating average precision (AP) separately for each class then divided by the average over the classes. Detection is reckoned as true positive if the mAP is above 0.5. As a result, the activity localization for each of the classes on 400 testing images is calculated by evaluating the overlap between a ground truth segment and a predicted temporal segment determined by Intersection over Union (IoU) score as shown in the section of training from scratch in Table 2 which summarizes all the prediction results from activity localization of RBN on 400 testing images by using YOLOv3 model trained from the collected dataset. We observe that the average precision of object classes of the runner is the highest, while that of object classes of the number is the lowest. This is due to several reasons, as shown in Figure 12.

First reason is the light condition, as some of the testing images are captured under insufficient light condition, because the particular marathon event is held from afternoon until evening and hence results in that the quality of images captured is not so good due to some blurriness. Secondly is that there are wrinkles in the racing bib that runners wear from images captured during approaching the finish line, since runners already ran for hours, which affects the result of RBN detection. Thirdly is that some regions of racing bibs are covered by the hands of the runner during running.

Other than mAP, there is another metric that is implemented for cascade network of YOLOv3 and CRNN models, which is the edit distance. Edit distance is a method of quantifying how dissimilar two strings are by calculating the minimum number of processes needed to convert one string into the other. There are three processes that can be employed to convert the string, which are: insert a character, remove a character and substitute a character. For instance, the edit distance of “2222” and “2221” is 1. Edit distance is generally employed in the field of natural language processing. For instance, automatic spelling correction can evaluate the candidate corrections for a misspelled word. This is done by choosing a word from a dictionary that has a minimum distance to the word by inquiry.

By referring to Figure 12, the graph illustrates the result of RBN detection and recognition of cascaded network before and after the improvement by fine-tuning with 1000 images. Edit distance is calculated for 400 testing images randomly selected from the testing subset of the first dataset. There are 64.25% of correct localization and recognition of RBN which is identical as ground truth numbers; 26.25% achieves one-number difference; 6% achieves two number-difference, 0.0075% achieves three number-difference and 2.75% achieves four number-difference or is completely different from ground truth numbers before improvement by transfer learning. Apart from that, results of edit distance that did not achieve zero edit distance or were not exactly the same as ground truth numbers have been studied and analyzed. There is a total amount of 143 images that did not attain zero edit distance. By referring to Figure 12, the reason of inaccurate text detector has achieved the highest percentage of 71.33%. In this work, YOLOv3 is utilized to localize the RBN from input images. However, since the mean average precision of ‘number’ obtains only 0.6224, it eventually affects the result of RBN recognition. Furthermore, there are other reasons that affect the result of RBN recognition as well, such as wrinkles of racing bib, mix-up between 1 and 7 and between 3 and 9, blurriness of input images captured among others, as illustrated in Figure 9. Although the ground truth number is ‘4203’, the result predicted from CRNN is ‘14203’. This is because of that the position of RBN is slightly rotated in the input image; hence patterns including logos or sketching will be detected by YOLOv3 model. Therefore, this results in that cropped image not only consists of RBN and subsequently affects the result of recognition of CRNN model.

There are some of the images which are still inaccurately recognized due to that the average precision of class of ‘number’ has achieved the lowest value which is only 0.6224. Therefore, the pre-trained model from YOLOv3 is fine-tuned with 1000 images by using transfer learning method, as shown in the section of fine-tuning in Table 3. Such 1000 images are collected from the same source with images from the first dataset used to train YOLOv3 from scratch. Transfer learning is known as a method of machine learning, where a model is implemented for a certain function again as the starting point for such model on another function or new problem. The priorities of transfer learning include saving time needed to develop and reutilizing an already developed model to do training for development on a different task or improvement. By referring to Table 2, it presents the mean average precision (mAP) before and after fine-tuning with 1000 training images captured from marathon running races. Results of localization of target classes about previous and current mAP values are tested on the same 400 training images. It can be observed that the localization accuracy of YOLOv3 is improved after transfer learning. Figure 13 illustrates the result of RBN detection and recognition of the cascaded network composed of YOLOv3 and CRNN before and after improvement by fine-tuning with 1000 images. Although YOLOv3 still achieved the highest result, there is a difference of 26 testing images that obtained accurate RBN recognition with 0 edit distance after transfer learning. Frequencies of other factors for inaccurate RBN recognition still remain unvaried, since the condition of wrinkles of racing bib and blurriness of testing images cannot be avoided.

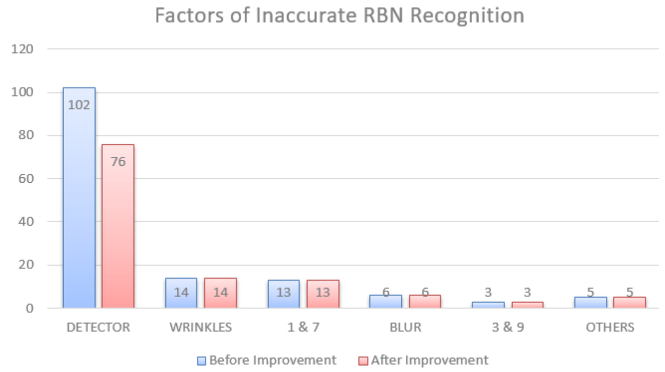


Figure 12. Factors of inaccurate RBN recognition before and after fine-tuning.

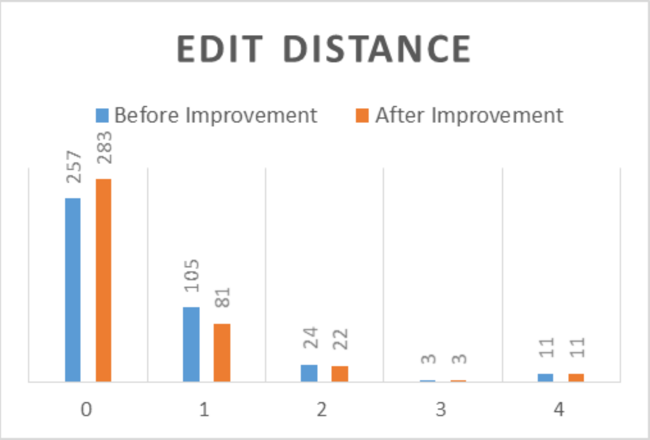


Figure 13. Edit distance before and after fine-tuning.

4. Conclusions

In running races, numerous images of runners are captured by running race organizers. This results in that the task of differentiating individual marathon images of a certain runner from all images becomes very troublesome. As a result, a deep learning-based RBN localization and recognition model for marathon running races is proposed. This is the first RBN recognition system implemented by using deep learning. Network topology implemented in this application is cascaded network composed of YOLOv3 and CRNN. YOLOv3 has been used for detecting the runner and bib number by multiple bounding boxes. CRNN is used to recognize the bib number through generating a label sequence for each input image and then selecting the label sequence that has the highest probability. Mean Average Precision (mAP) achieves 85.15% accuracy and the edit distance with exact output is 283 out of 400 after fine-tuning. The recognition could not reach 100% of accuracy because of that some RBNs in input images are in the condition of wrinkles or patterns such that sketching or logo surroundings around RBN are accidentally predicted together with RBN hence affecting the accuracy of cascaded network. Accuracy of detection has been further improved through transfer learning by fine-tuning with a validation dataset which consists of 1000 images. 9% to 21% of improvement to the accuracy has been achieved. With the demonstrated capability of automating the localization and recognition of racers, the developed system not only could better control large-scale marathon events, but also tighten the security of racing events.

ACKNOWLEDGEMENTS

The authors acknowledge the technical and financial support by Universiti Teknikal Malaysia Melaka (UTeM).

REFERENCES

[1] S. Roy, P. Shivakumara, P. Mondal, R. Raghavendra, U. Pal and T. Lu, "A New Multi-modal Technique for Bib Number/Text Detection in Natural Images," Proceedings of Pacific Rim Conference on Multimedia, pp. 483-494, 2015.

[2] K. S. Younis, "Arabic Handwritten Character Recognition Based on Deep Convolutional Neural Networks," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 3, no. 3, pp. 186-200, 2017.

[3] P. Shivakumara, R. Raghavendra, L. Qin, K. B. Raja and T. Lu, "A New Multi-modal Approach to Bib Number / Text Detection and Recognition in Marathon Images," Pattern Recognition, vol. 61, pp. 479–491, 2017.

[4] T. Basha, S. Avidan and I. Ben-Ami, "Racing Bib Number Recognition," Br. Mach. Vis. Conf. (BMVC), pp. 1–10, 2012.

[5] B. Epshtein, "Detecting Text in Natural Scenes with Stroke Width Transform," Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2963–2970, 2010.

[6] F. Su and H. Xu, "Robust Seed-based Stroke Width Transform for Text Detection in Natural Images," Proc. of the 13th Int. Conf. Doc. Anal. Recognit., pp. 916–920, 2015.

[7] R. Smith, "An Overview of the Tesseract OCR Engine," Proc. of the 9th International Conference on Document Analysis and Recognition, pp. 629-633, 2005

[8] N. Boonsim, "Racing Bib Number Localization on Complex Backgrounds," WSEAS Transactions on Systems and Control, vol. 13, pp. 226–231, 2018.

[9] Q. Ye and D. Doermann, "Text Detection and Recognition in Imagery : A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 7, pp. 1480–1500, 2015.

[10] Y. Zheng and A. R. O. Theory, "Edge Detection Methods in Digital Image Processing," Proc. of the 5th Int. Conf. Comput. Sci. Educ., pp. 471-473, 2010.

[11] Celsius, "Delaware Running Festival 2018," [Online], Available: https://pic2go.nascent-works.com/c0ab85474c3597a8dcd1a3d95e917516, [Accessed: 15-Apr-2019].

[12] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 779-788, 2016.

[13] J. Redmon, A. Farhadi and C. Ap, "YOLOv3 : An Incremental Improvement," [Online], Available: https://arxiv.org/abs/1804.02767, 2018.

[14] OpenCV, "AI Courses by OpenCV," [Online], Available: https://opencv.org/ [Accessed: 15-Apr-2019]

[15] B. Shi, X. Bai and C. Yao, "An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2298–2304, 2016

[16] N. Aloysius, "A Review on Deep Convolutional Neural Networks," Proc. of Int. Conf. Commun. Signal Process., pp. 588–592, 2017.

[17] H. Li, P. Wang and C. Shen, "Towards End-to-end Text Spotting with Convolutional Recurrent Neural Networks," IEEE Int. Conf. Comput. Vis., no. 2, pp. 5248–5256, 2017.

[18] A. Graves, M. Liwicki, S. Ferna, R. Bertolami and H. Bunke, "A Novel Connectionist System for Unconstrained Handwriting Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 5, pp. 855–868, 2009.

[19] A. Graves and S. Fern, "Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks," Proceedings of the 23rd International Conference on Machine Learning, pp. 369-376, 2006.

[20] Y. C. Wong and Y. Q. Lee, "Design and Development of Deep Learning Convolutional Neural Network on a Field Programmable Gate Array," Journal of Telecommunication, Electronic and

Computer Engineering, vol. 10, no. 4, pp. 25-29, 2018.

[21] T. Vo, T. Nguyen and C. T. Le, "Race Recognition Using Deep Convolutional Neural Network," Symmetry, vol. 10, no. 11, 564, 2018.

ملخص البحث:

يتزايد الاهتمام بأنماط الحياة الصحية في جميع أنحاء العالم. وقد جرى تنظيم أعداد كبيرة من سباقات الماراثون التي استقطبت اهتمام الكثيرين من الناس من مختلف الأعمار من الذكور والإناث من بلدان مختلفة. وما من شك في أن هذا التنوع الواسع يصعب استيعاب الأعداد الضخمة من الصور في سباقات الماراثون؛ نظراً لأن هذه العملية تتم يدوياً. لذا، تقترح هذه الورقة نهجاً جديداًلكشف أرقام المتسابقين في سباقات الماراثون وتمييزها باستخدام التعلم العميق. وفي السابق، كان ذلك يتم باستخدام تقنيات معالجة الصور فقط؛ الأمر الذي يحدّ من جودة الأداء.

يتألف النظام المقترح من مرحلتين هما: كشف أرقام المتسابقين، وتمييز تلك الأرقام. في المرحلة الأولى، تستخدم شبكة التفافية (YOLOv3) مفردة لتوقُّع المتسابق ورقم المتسابق من خلال صناديق تحديد متعددة الى جانب الاحتمالات المتعلقة بصنوف تلك الصناديق. والجدير بالذكر أن تلك الشبكة هي شبكة تصنيف تفوق في أدائها مثيلاتها من الشبكات المستخدمة. أما في المرحلة الثانية، فيجري استخدام شبكة التفافية أخرى (CRNN) لتوليد تتابع من العلامات لكل صورة مُدخلة ومن ثم اختيار تتابع العلامات الذي يحظى بالاحتمال الأعلى. ويمكن تدريب تلك الشبكة مباشرة من العلامات المتتابعة مثل الكلمات غير المحتوية على حواشٍ تفسيرية. وبذلك تقوم شبكة (CRNN) بتمييز محتويات أرقام المتسابقين التي تم كشفها. من ناحية أخرى، جرى تحليل كل من متوسط الدقة ومسافة التحرير. واتضح أن النظام المطّور مناسب لأحداثٍ مثل سباقات المسافات الطويلة ومنها سباقات الماراثون؛ فهو يعمل على أتمتة عملية تحديد موقع المتسابق وتمييزه، ومن ثم فهو يزيد من فاعلية السيطرة على السباقات ومراقبتها، الى جانب تحسين المعالجة البَعدية للبيانات الخاصة بتلك الأحداث.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 05, No. 03, December 2019.

MULTI-LEVEL ANALYSIS OF POLITICAL SENTIMENTS USING TWITTER DATA: A CASE STUDY OF THE PALESTINIAN-ISRAELI CONFLICT

Iyad Al-Agha and Osama Abu-Dahrooj

(Received: 9-Jul.-2019, Revised: 15-Aug.-2019, Accepted: 3-Sep.-2019)

ABSTRACT

Given the tremendous growth of social media platforms, people have been actively spreading not only information in general, but also political opinions. Many research efforts have used social media content to analyse and predict the public opinion towards political events. This work presents an analytical study for measuring the political public opinion towards the Palestinian-Israeli conflict by using Twitter data. The study uses a novel data analysis model that leverages two levels of analysis: country-level analysis and individual-level analysis. The country-level analysis aims to explore the country's overall attitude towards Palestine by: 1) Identifying counties that generated the most topic-focused tweets, 2) Measuring the friendliness of each country towards Palestine. 3) Analysing the change of sentiment over time. The individual-level analysis aims to analyse data based on the activity and background of individuals. The attitudes of opinion leaders and ethnic groups were analysed and discussed in light of countries' attitudes.

The rich experience provided in this study through the proposed model for multi-level analysis, the step-by-step procedure, the variety of analysis techniques and the discussion of results can be informative for other developers and data analysts who are interested in analysing social media sentiment about political conflicts in particular.

KEYWORDS

Sentiment analysis, Public opinion, Twitter, Politics, Palestinian-Israeli conflict.

1. INTRODUCTION

With the dramatic rise of social media in the past decade, millions of people express their views on a great variety of topics. This has dramatically increased the data available to mine social media platforms, such as Twitter, for information about how people think and feel. Identifying the public opinion towards political issues is essential to shape the international policies, alliances and positions. Governmental officials should pay attention to public opinion to decide how to act. Opinion polls have been the standard mechanism for collecting public opinion. However, polls have several problems that have been reported in the literature [1]-[2]. They often use samples of small sizes or non-representative samples that may result in inaccurate findings. Unclear, biased or emotionally charged questions will produce misleading answers and weaken the accuracy of the results of a poll. In addition, the results of opinion polls are perspective as their findings apply only at the time the questions were asked. However, the public opinion towards a particular issue is likely to fluctuate over time based on recent updates. Finally, with polls it is difficult to perform fine-gained analysis or to understand the subjectivity and the motivations behind the public opinion. All the aforementioned limitations make opinion polls not very reliable and there is a need for other mechanisms to capture the public opinions.

Social media platforms have grown explosively over the past decade. People from all over the world have been using them extensively to express their views and discuss topics of interest. The large number of users, the variety of discussed topics and the massive volumes of posted content have made social media a rich source to understand and predict the population attitudes. Mining social media for political opinions may provide a faster and less expensive alternative to traditional polls.

Numerous research works have explored the mining of social media to analyse or predict political opinions [3]-[5]. However, these works were mostly event-specific and used techniques relevant only to the issue being investigated. In addition, most studies relied on sentiment analysis that primarily aims to predict the user feeling rather than the political opinion. In politics, judging a sentiment depends on

I. Al-Agha and O. Abu-Dahrooj are with Department of Computer Science, Faculty of Information Technology, The Islamic University of Gaza, Palestine. Emails: ialagha@iugaza.edu.ps and osama.abudhrooj@gmail.com

the side you stand by regardless of the user's emotions. Thus, the same idiom may be interpreted differently based on the context. For example, the idiom: "I feel sorry for the Palestinian people☹" conveys a feeling of sadness and sympathy, thus may be classified as "negative" by a conventional sentiment analyser, despite that it carries a positive attitude towards the Palestinian case. Similarly, expressions that evoke a positive emotion towards Israel, such as "I love Israel" should have a negative polarity from the perspective of Palestine. These examples show that the conventional sentiment analysis that is based on feelings or emotions may be inadequate for inferring political attitudes that are based on a specific understanding of what is "positive" and what is "negative". In addition, few studies [6]-[7] have used social media to explore the public opinion towards the Palestinian-Israeli conflict and the majority have used statistics to analyse existing situations rather than making predictions about the public attitudes.

In this paper, we propose an analytical study that uses a sample of Twitter text data in English to measure and analyse the political public opinion in several countries around the world towards the Palestinian-Israeli conflict. The proposed approach builds on a data analysis model that leverages two levels of analysis: country-level analysis and individual-level analysis. Several types of analysis are used under each level, each of which aims to provide an insight into a particular aspect. The aim is to provide more in-depth analysis that leads to a better understanding of the public opinion. This work demonstrates, through a realistic case study and a step-by-step procedure, how different data mining techniques, such as sentiment analysis, time-based analysis and opinion leaders analysis can be used to gain deeper and more fine-grained insights. We believe that the proposed analysis model can be adapted and reused for similar studies, especially those focusing on sentiment analysis of social media content about political conflicts.

2. RELATED WORKS

In the past decade, a growing number of studies have used data from Twitter to monitor sentiments for the purpose of tracking trends in politics, economy and public opinion [8]-[9]. In general, these studies can be classified into three research areas based on the purpose they used sentiment analysis for [1], [10]. The first research area concentrates on predicting real-world continuous values by analysing sentiments in social media, such as predicting stock market values. Several studies in this area have reported a notable improvement in prediction results when incorporating sentiments [11]-[13]. It is noteworthy that the focus in these studies was on emotive sentiment; i.e., mood states, rather than on polar sentiment (positivity, negativity) which is popular in politics.

The second area is result forecasting. A popular example is the predication of election results. Twitter has been used increasingly in the past decade to forecast the public opinion in the events of elections [3], [14]-[16]. Researchers in this area have focused either on determining current levels of support toward political actors or on predicting support in upcoming elections. However, they often had different opinions about the reliability of social media mining as a prediction tool in the time of elections. Some studies tried to validate sentiment measured from Twitter by comparing it with the public opinion measured from polls [4] or by comparing it with the final election results [1]. These studies reported high correlations between actual and predicted results and confirmed the potential of social media in predicting political views. In contrast, other studies underestimated the prediction based on social media due to many flaws, such as the untrustworthy content, the negligence of demographics, the non-representative sample and the inaccurate ways used to validate results [15], [17] . Overall, studies from both sides have supported the use of social media mining as a supplement for traditional polling.

The third related area is event monitoring, where the aim is to analyse the social media reactions to specific events. These events could be election debates [18]-[20], campaigns [3], [21] or any political event affecting the public opinion [7], [22]. Many works in this area focused merely on using sentiments of tweets to understand their relationships to the event of interest [3], [23]. Other works tried to enhance the sentiment analysis by combining it with other factors, such as the retweet behaviour [24], hashtags [25], the opinion leaders' behaviour [26]-[27] and contextual information, such as geo-location, temporal and author information [28]-[30]. In general, the previous studies concluded that Twitter proved to be an effective source of data for monitoring and assessing the public reaction associated with important events.

The work in this paper falls under the third area, as it aims to characterize the international public attitude towards the Palestinian-Israel conflict in terms of Twitter sentiment. It builds on the methods used in the literature and contributes in providing in-depth tracking of public opinion at multiple levels: the country level and the individual level.

When it comes to the Palestinian-Israeli issue, very few studies have exploited social media content to capture patterns or trends related to the ongoing conflict [6], [7], [31]. These studies, however, relied solely on systematic statistical reviews rather than on data mining or sentiment analysis. In addition, they were published in the domain of politics, where the focus was on the findings and implications rather than on the underlying technology. Although our findings may complement those of previous studies, the emphasis of this work is on describing the used approach and techniques.

Several efforts showed an interest in the online sentiment analysis to predict the result of elections and monitor political events. Tumasjan, Sprenger, Sandner and Welpe [2] presented the first attempt to investigate whether Twitter validly mirrors the German election results. They analysed about 100,000 political tweets identifying either a politician or a political party and used LIWC2007 [32] tool for extracting sentiment from the tweets. Burnap, Gibson, Sloan, Southern and Williams [33] used Twitter to forecast the outcome of 2015 UK general elections. They used an approach that incorporates sentiment analysis and prior party support to generate a forecast of parliament seat allocation. Ceron, Curini and Iacus [34] and Ceron, Curini and Iacus [35] attempted to analyse several elections in Italy, France and the US and showed that a supervised learning method developed by Hopkins and King [36] does a good job of explaining fluctuation in party or candidate support in various contexts. However, the previous works put emphasis on using emotional states to identify user preferences and did not examine the influence of other factors, such as individual characteristics, geo-locations and opinion leaders.

Besides the sentiment analysis of Twitter data, some efforts tried to incorporate other sources of user-generated media. For example, Le, Boynton, Mejova, Shafiq and Srinivasan [37] studied Twitter communications around the 2016 U.S. elections and implemented computational methods for tracking political discourse about party, personality traits and policy on Twitter. O'Connor, Balasubramanyan, Routledge and Smith [4] investigated the people’s remark measured from polls with opinion measured from microblogging sites. They used time series to assess the population's aggregate opinion on a topic and measured correlations to several polls conducted during the same period of time. Marozzo and Bessi [38] presented a study analysing the polarization of social network users and news sites during political campaigns characterized by the rivalry of different factions. They performed temporal analysis to monitor the changes of polarization during the weeks preceding the vote. Martin-Gutierrez, Losada and Benito [39] analysed temporal series and interaction networks corresponding to two Twitter datasets downloaded during the Spanish electoral campaigns of 2015 and 2016 in order to identify recurrent patterns of user behaviour. Cody et al. [40] studied the sentiment surrounding climate change conversation on Twitter and used temporal analysis to observe how sentiment varies in response to climate change news and events. Our work also extends the sentiment analysis by using temporal analysis and incorporating multiple factors. It also analyses the influence of each factor on the overall public opinion. However, it has a different objective as it focuses on the Palestine-Israeli issue.

The contributions of this work reside in the following: First, it proposes a multi-level model of analysis that leverages multiple factors at both group and individual levels and employs computational methods to perform fine-grained analysis of public opinion. This is different from most of the existing efforts that focused solely on sentiment analysis or used a certain type of features. Second, the work discusses the special considerations for political sentiment analysis and demonstrates, through an experimental study, that sentiment analyses that are based on emotional states may be inadequate for inferring political polarization. We believe that the detailed procedure and computational methods reported in this work can be informative to data analysts and practitioners in investigating political conflicts in particular.

3. Overview of The Proposed Approach

The overall approach used to undertake this study is depicted in Figure 1. It consists of the following steps: data collection, data pre-processing, political sentiment analysis and feature extraction and analysis. The following sections start by describing the data analysis model consisting of two levels of analysis: country-level analysis and individual-level analysis. The features that should be extracted to realize the proposed model are also described.

Afterwards, a case study that utilizes the proposed model to analyse the international public opinion towards the Palestinian-Israeli conflict is presented in detail. Data collection and pre-processing steps are described, highlighting the considerations we took to assure that the collected data will not lead to invalid or biased results. At the core of our approach resides the sentiment analysis step. The paper reports on the experiments conducted to compare several sentiment classifiers and to train and evaluate our own sentiment classifier.

The sentiment analysis step is followed by feature extraction, in which several features are extracted or derived from the inferred sentiments. Extracted features are then used to carry out the analysis at both country and individual levels. Finally, results are presented and discussed.

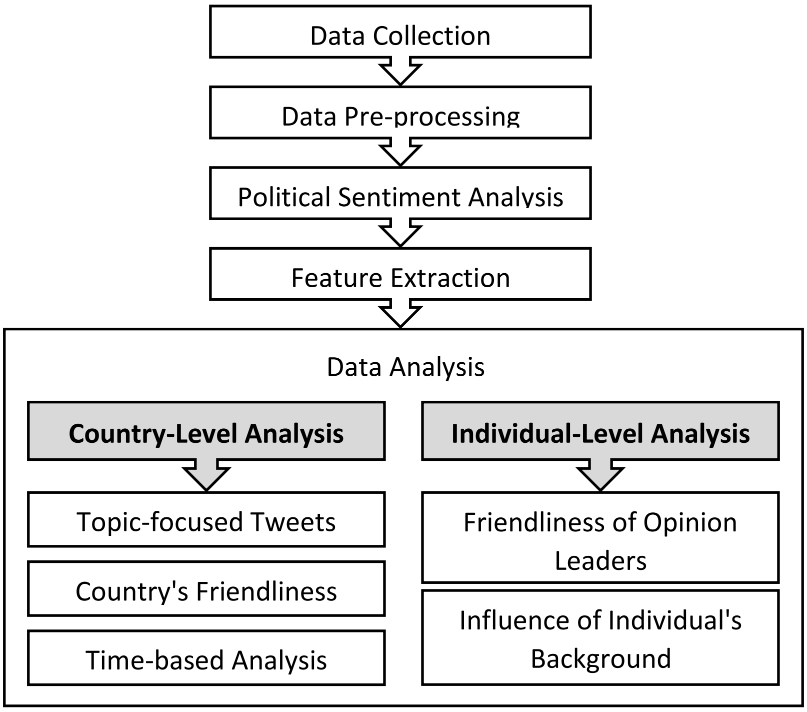


Figure 1. Approach for analysing public opinion towards the Palestinian-Israeli conflict.

4. Data Analysis Model

The data analysis model consists of two levels of analysis as shown in Figure 1. These levels are explained as follows:

4.1 Country-level Analysis

The purpose of the country-level analysis is to explore the country's overall interest in and attitude towards the political issue being studied. Country-level analysis is done through the following:

- Identifying counties that generated the most tweets related to the political issue; i.e., topic-focused tweets. The aim in our case study is to determine countries that show, on Twitter, the most concern about and awareness of the Palestinian-Israeli issue regardless of sentiment. Thus, tweets are counted per country, while sentiment scores are ignored.
- Measuring the friendliness of each country towards Palestine: Friendliness of a country indicates the level of support and sympathy it shows towards one side and can be determined from the polarities of tweets. In this study, friendliness is defined from the perspective of Palestine, so that positive and supportive attitudes towards Palestine lead to high friendliness rates. In contrast, views opposing the Palestinian side or advocating for the opposite side yield low friendliness rates.
- Analysing data across time to investigate how the public opinion changes over time.

4.2 Individual-level Analysis

The individual-level analysis aims to analyse data based on the activity of individuals and their backgrounds. This is performed through the following:

- Capturing the attitudes of opinion leaders. The term "opinion leader" refers to an active user on social media who has a large number of followers and can influence the opinions and behaviours of others [41]-[42]. Identifying opinion leaders is crucial to promote behaviour change or to identify subjects that are of high interest to people [43]. Measuring the attitude of opinion leaders towards political issues is important, because they reflect large sectors in their communities.
- Capturing the individual's characteristics: The individual's background or characteristics, such as nationality, religion, ethnicity and gender, may influence his/her political stance. For example, women are likely to stand in favour of issues pertaining to women rights and Arabs and Muslims are more keen to support the Palestinian rights. Identifying these characteristics from social media, where possible, will help better understand the motivations behind the public opinion. However, deciding which characteristics to capture and analyse is case-specific and depends on the objectives of the case study. For example, this work sought to measure the influence of the individual's ethnicity on the public opinion and the potential relationship between the ethnicity of users and their perceptions of Palestine.

5. EXTRACTED FEATURES

To achieve the data analysis model as explained above, several features need to be extracted from the collected tweets. These features are as follows:

- Polarity**:** Polarity is the sentiment score of the tweet, which determines the classification of the tweet (e.g. positive, negative or neutral). Polarities of tweets are measured by using the sentiment analyser. Other features will be derived from the polarities of tweets.
- Friendliness**:** The friendliness of a country is measured by calculating the average polarity of tweets posted by users in the country. Similarly, friendliness of an individual is the average polarity of tweets posted by the individual. To compute the average polarity, we interpreted the three sentiment values: positive, neutral and negative into +1, 0 and -1, respectively. Then, the friendliness for a country F_c is computed using the following equation:

$$\boldsymbol{F_c = \frac{\sum Polarity(t_i)}{n} \times 100} \qquad (1)$$

  where, *n* is the number of tweets attributed to the country *c*. t_i is a tweet posted from the country *c*. The friendliness score ranges between -100 and +100, where +100 denotes the maximum friendliness value.
- Leadership**:** This feature is used to identify opinion leaders. Different metrics have been used in the literature to identify opinion leaders [42], [44]. In this work, Twitter users in each country who have the most number of followers are treated as opinion leaders.
- Individual's characteristics**:** In this study, the aim is to identify the individual's ethnicity from the user's name or nickname and then to analyse the influence of inferred ethnicities on the public opinion.

6. DATA COLLECTION

Twitter's public API is a streaming API offered by Twitter for collecting tweets. Although it has been widely used in the literature, it has a drawback in that it provides only 1% or less of its entire traffic, without control over the sampling procedure, which is likely insufficient for accurate analysis of public sentiment [45]. Instead, we used a Twitter search analytics and business intelligence tool called Followthehashtag [46]. Followthehashtag enables searching for tweets over a specific period of time. We first used Google Trends(1) to find top search keywords used in Palestine and Israel that are related to the Palestinian-Israeli conflict over the year 2016. Then, we selected keywords that represent the

(1) Google Trends, https://trends.google.com/trends/?hl=en

opposite views of the two sides of the conflict in order to avoid biased results. Examples of selected keywords include: Palestinian-Israeli conflict, Israeli occupation, Apartheid wall, settlements, Gaza, West bank, Judea and Samaria, Jerusalem, Palestinian terrorism and suicide bombings. Finally, we used these keywords to perform a query-based search to collect tweets related to the conflict that were posted during the year 2016.

In total, 178,524 tweets were collected. These tweets were posted by approximately 48,531 users during the period from Dec. 20 2015 to Dec. 31 2016. We think a period of one year is sufficient to explore the political trends on Twitter and to perform time-based analysis, since many related studies used equal or shorter periods (e.g. [1, 2, 5, 31]).

The following information was retrieved for each tweet: the tweet's text, the username and nickname of Twitter user, date and time of posting , country and place of origin, number of followers of the tweet's author, number of users followed by the author and hashtags. Most of these tweets were from the US, UK, Canada, Australia, Finland and some other European countries. 89.78% of the collected tweets were in English. Table 1 shows statistics about the collected tweets. The whole dataset can be found on the following link: https://github.com/odahroug2010/2017.

Table 1. Statistics about collected tweets.

General information	Total# of tweets	178,524
	Number of users	48,530
	Duration	Dec. 20 2015 to Dec. 31 2016
	English tweets	89.78%
	Retweets	7948
	Avg. no. of words per tweet	12.74
	Standard Dev. of words	5.002
Location information	No. of countries	174
	Top sources of tweets	US, UK, Canada, Australia, Finland and other European countries
	No. of tweets with unknown sources	28156
	No. of retweets	7948
	Min. tweets by country	24
	Max. tweets by country	27490
	Avg. tweets by country	777.88
	Standard Deviation	3363.68

7. DATA PRE-PROCESSING

Tweets often have special characteristics that make their pre-processing different from that of ordinary texts. Tweets are of limited length (140 characters at most) and may contain special texts, such as hashtags, URLs, emoticons and usernames. For the pre-processing of tweets, we used the approach depicted in Figure 2, which consists of the following steps:

- Filtering: Collected tweets were filtered by: 1) removing non-English tweets: 10.22% of collected tweets were written in non-English languages and thus were excluded, 2) removing tweets with unknown resources: 28156 tweets in total did not have countries of origin. These tweets were excluded, because they are out of the scope of our analysis, 3) removing re-tweets: 7948 of tweets were retweeted and these were excluded from the dataset, so that only original tweets are counted. 124,174 tweets remained after the filtering step.

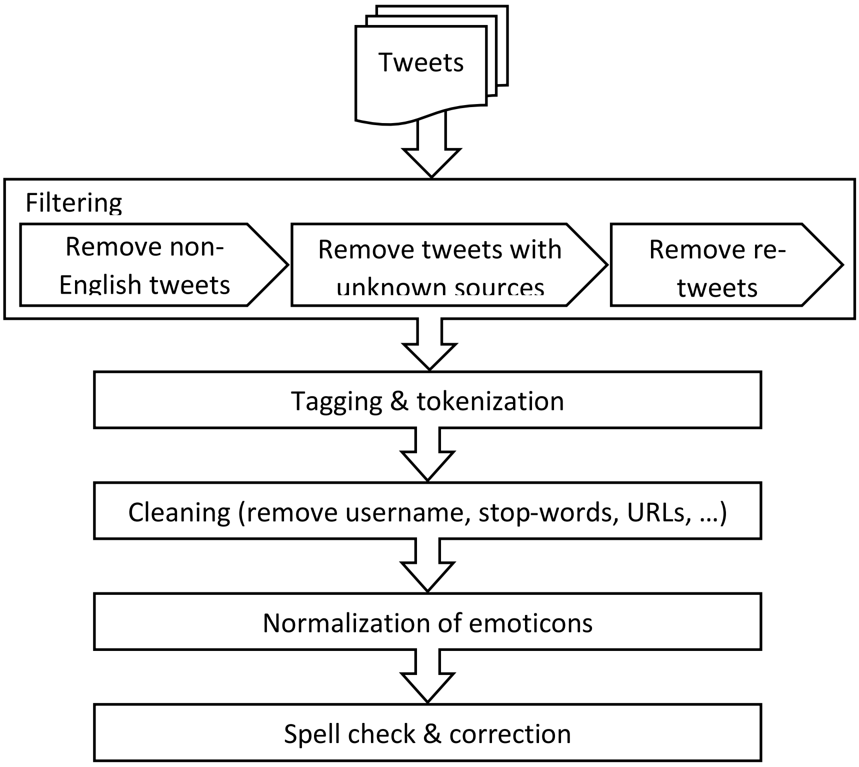

Tweets
Filtering
Remove non-English tweets
Remove tweets with unknown sources
Remove re-tweets
Tagging & tokenization
Cleaning (remove username, stop-words, URLs, …)
Normalization of emoticons
Spell check & correction

Figure 2. Pre-processing steps of tweets.

- Tokenization and tagging**:** As Twitter allows users to write short texts only, tweets often come with a special grammar and abbreviations, so that users can convey the messages with least possible words. Traditional tokenizers and POS taggers may be inadequate for pre-processing tweets and there is a need for alternatives that can recognize tweet's tokens, hashtags, emoticons and URLs. We used a text processing library called ArkTweetNLP to tokenize and tag tweets [47]. The library was developed specifically to handle informal and online conversational text including various non-standard lexical items and syntactic patterns.
- Cleaning: Twitter users prefer to use symbols and non-standard language in their tweets. Many of the used symbols may be irrelevant and thus should be excluded to avoid an incorrect result when applying the sentiment analyser. In our approach, tweets were cleaned by removing the following parts: usernames, numeric expressions, punctuations, URLs and stop-words that are unlikely to affect sentiments. These parts were recognized from the tagger applied in the previous step.
- Normalization of emoticons: Emoticons are important for sentiment analysis; thus, their meanings should be preserved and they should not be removed from the tweets. In our approach, we used a special dictionary that contains the most used emoticons and their meanings in English [48]. This dictionary was used to replace each emoticon with its relevant meaning. These examples show that conventional sentiment analysis that is based on feelings or emotions may be inadequate for inferring political attitudes that are based on a specific understanding of what is "positive" and what is "negative".
- Spell check and correction: Tweets may contain incorrect or miss-spelled words and this will affect the result of sentiment analysis. This step manipulates these words by using a spell checker and substitutes them with correct words. Jazzy Spell Checker [49] was used for this step. As an example of the output of this step, a tweet like "I looove palestin. Happi to visit it" will be corrected to "I love Palestine. Happy to visit it".

8. Sentiment Analysis

Sentiment analysis is the core step to identify attitudes towards the Palestinian-Israeli conflict. When sentiments are identified, tweets can be categorized based on different features. Therefore, the results of subsequent steps largely depend on the quality of sentiment analysis. It is assumed that the tweet is an opinion and therefore we need to know its polarity classification, which is positive, negative or neutral. To achieve this, we used a supervised approach for sentiment analysis.

As mentioned earlier, it is important to emphasize that the sentiment analysis in this work aims to identify the political stance rather than mere the user feeling. In political conflicts, as in our case study, the polarity of a tweet should be determined based on the side you stand by regardless of the expressed emotions. As an example, tweets that show support to Israel are assessed as 'negative' from the perspective of Palestine even if they convey positive emotions. Although there are several pre-trained "off-the-shield" tools to perform sentiment analysis, these tools are often trained to identify feelings or emotions rather than political sentiment and thus they may be inadequate for the purpose of this study. Therefore, we decided to build our own sentiment classifier by training it on a manually-labelled dataset. Then, the performance of the classifier will be assessed by comparing it with other pre-trained sentiment analysers.

Since the collected tweets do not come with predefined sentiments, we decided to pick a sample of tweets and label them manually with the relevant polarity (positive, negative or neutral). These labelled tweets will be then used to train and evaluate the sentiment analyser. 1300 tweets (about 10% of the entire dataset after the filtering step) were chosen randomly and given to two human subjects to label them separately. In general, the labelling of tweets was done from the perspective of Palestine based on the following criteria:

- Tweets that include appreciation, praise, glorification or support for Palestine or the Palestinian issue were labelled as positive. For example, idioms like "Free Palestine" or "It is called Palestine, not Israel" are assigned positive polarity.
- Tweets that show solidarity and sympathy with Palestine or Palestinians were labelled as positive. For example, idioms like "Please donate for the children of Gaza" or "Save Palestinian children …" should be labelled as positive.
- Tweets that contain idioms denoting negative attitude towards "Israel", e.g. "Stop the Israeli apartheid wall" are considered positive from the perspective of the pro-Palestinian point of view.
- Tweets that show clear support for or sympathy with "Israel" were labelled as negative. For example, idioms like "I love Israel" or "Israel has the right to defend itself" all carry positive attitude towards "Israel" and negative attitude towards Palestine and thus were labelled as negative.
- Tweets that use Israeli naming conventions, such as "Judea and Samaria", "IDF army" and "Palestinian terrorists" were treated as negative sentiments, since they adopt a pro-Israel stance.

After analysing labels received from the two subjects and ignoring disagreements, we ended up with 1264 tweets, of which 637 were positive, 543 were negative and 84 were neutral.

Sentiment analysis in this work was carried out using a logistic regression model implemented by LingPipe [50]. LingPipe classifies texts by using a language model on character sequences and the execution uses the 8-gram language model. The labelled 1264 tweets were randomly split into two parts: 80% of the tweets (1011 tweets) were used for training and 20% (253 tweets) were used for testing. 10-fold cross validation was performed.

Table 2 shows the testing results of the trained classifier. Precision and recall values for each class were calculated by creating the confusion matrix. The matrix shows that the classifier achieved good results with positive and negative tweets, but the performance was low with neutral tweets. However, the low performance in case of neutral tweets will have a marginal impact on the results due to the low number of neutral tweets in general.

The performance of our sentiment classifier was also evaluated by comparing it with other pre-trained sentiment classifiers that are: Stanford CoreNLP [51], SentiStrength [52] and the pre-trained LingPipe. These classifiers were chosen, because they are frequently used to analyse the user sentiments on social media, especially in political events [53-58]. The testing dataset used above for testing our classifier was also used for testing the other classifiers.

The comparison results are shown in Table 3. Results indicate that our classifier outperformed the other classifiers significantly. It is also obvious that the performance of the pre-trained classifiers was remark-

Table 2. The confusion matrix of the trained sentiment classifier.

	Label Positive	Label Negative	Label Neutral	Total Predicted	Precision	Recall
Predict Positive	104	12	2	118	88.1%	80%
Predict Negative	18	93	3	114	81. 6%	83.8%
Predict Neutral	8	6	7	21	33.3%	58.3%
Total Label Class	130	111	12			

ably poor. This can be attributed to the fact that they are designed to infer polarity based on emotional states that often contradict with political attitudes. This proves that the traditional sentiment analysis may be inadequate for inferring political polarization, where the polarity becomes a relative issue depending on the perspective of the interpreter and the case being analysed.

Table 3. Comparison between sentiment classifiers.

S. No.	Classifier	Accuracy	Precision	Recall	F-measure
1	Stanford CoreNLP	8.1%	30.6%	22.6%	26%
2	SentiStrength	7.9%	42.2%	27.8%	33.5%
3	Pre-trained LingPipe	31.2%	35.6%	30.5%	32.9%
4	**Our classifier**	**80.63%**	**81.35%**	**80.61%**	**81%**

The sentiment classifier built was used to measure polarities of 169694 tweets; these are the whole collected tweets excluding the retweets and the tweets used for training and testing of the sentiment classifier.

9. Data Analysis and Results

The sentiment classifier built in the previous section was used to measure the sentiments of all tweets in the dataset, excluding those used to build and test the classifier. In total, the sentiments of 122,921 tweets were measured.

The following sub-sections describe the application of the proposed analysis model, see Figure 1, on measured sentiments in order to derive the features needed to analyse the international public opinion. Afterwards, the main findings are presented, discussed and validated where possible. Apache Spark [59], which is an analytics engine for large-scale data processing, was used to implement the analysis model.

9.1 Country-level Analysis

The country-level analysis includes three types of analysis: countries that generated the most topic-focused tweets, friendliness of countries and time-based analysis. Results of each analysis is explained as follows.

9.1.1 Countries that Generated the Most Topic-focused Tweets

The volume of tweets that can be attributed to each country was measured. At this stage, polarities of tweets were ignored and the focus was only on counting the number of tweets per country.

Each tweet in the dataset often comes with geo-information that help identify its country of origin. One attribute is called "country" and it should be set with the country code. For example, tweets posted from the UK have the country value "GB". However, the country code may be missing for many tweets and it can be identified only if it is set in the user profile. Tweets can also have geocoding attributes named "Latitude" and "Longitude". These attributes are set to valid values for tweets posted from devices with enabled GPS service. For tweets that have latitude-longitude values but the country value is missing, the Google maps geocoding service was used to determine the corresponding countries. After assigning tweets to countries, tweets were counted per country and countries that ended up with a number of tweets less than 0.1% of the total number of tweets were ignored.

Table 4 shows the top ten countries in terms of the number of tweets concerning the Palestinian-Israeli issue. Canada, the UK and the US generated the most tweets. This result is expected considering the high involvement of these countries in the Middle East affairs. The bottom countries were Slovenia, New Zealand and Austria. Figure 3 illustrates the results on a geographical map. When considering the number of population, Jersey, Canada and Finland generated the most tweets *per capita*. The bottom countries were Nigeria, India and China.

Table 4. Top ten countries in terms of the number of tweets related to the Palestinian-Israeli conflict.

No.	Country	Country code	Focused Tweets
1	Canada	CA	27,490
2	United Kingdom	GB	23,010
3	United States	US	20,125
5	Ecuador	EC	9,342
6	Finland	FI	3,654
7	Australia	AU	3,125
8	Netherlands	NL	2,646
9	India	IN	1,445
10	France	FR	1,215

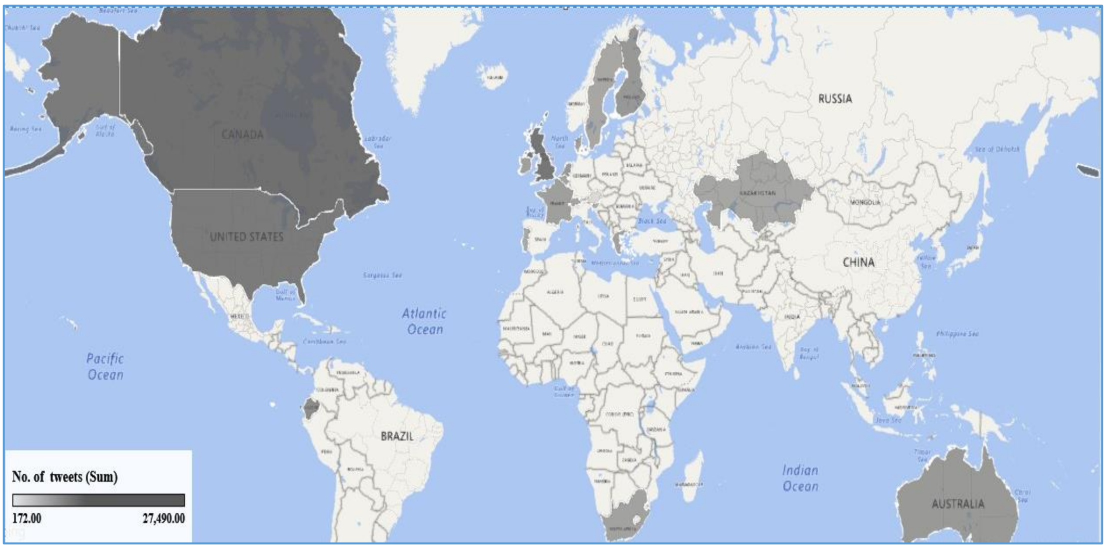


Figure 3. Number of tweets per country.

To get insight into the validity of the above results, we compared these results with the corresponding country indices generated from Google Trends. Google Trends provides a metric called Google index that indicates the frequency at which people in a country search for the term during a specific period of time. We used Google index to measure the frequency at which people search for the terms related to the Palestinian-Israeli issue in each country from the top 30 countries that posted the most tweets according to our results. The search activity was measured from January 2016 to December 2016, which is the same period through which the tweets were collected. Our assumption is that the search activity, measured through Google index, should be consistent with the tweeting activity during the same period of time. To make easy comparison, the tweet counts were normalized by log-transform, so that they become comparable with values from Google index (The log of tweet count is named as Twitter index) [60]. We plot the Google index as the x-axis and the Twitter index as the y-axis. The result is depicted in Figure 4.

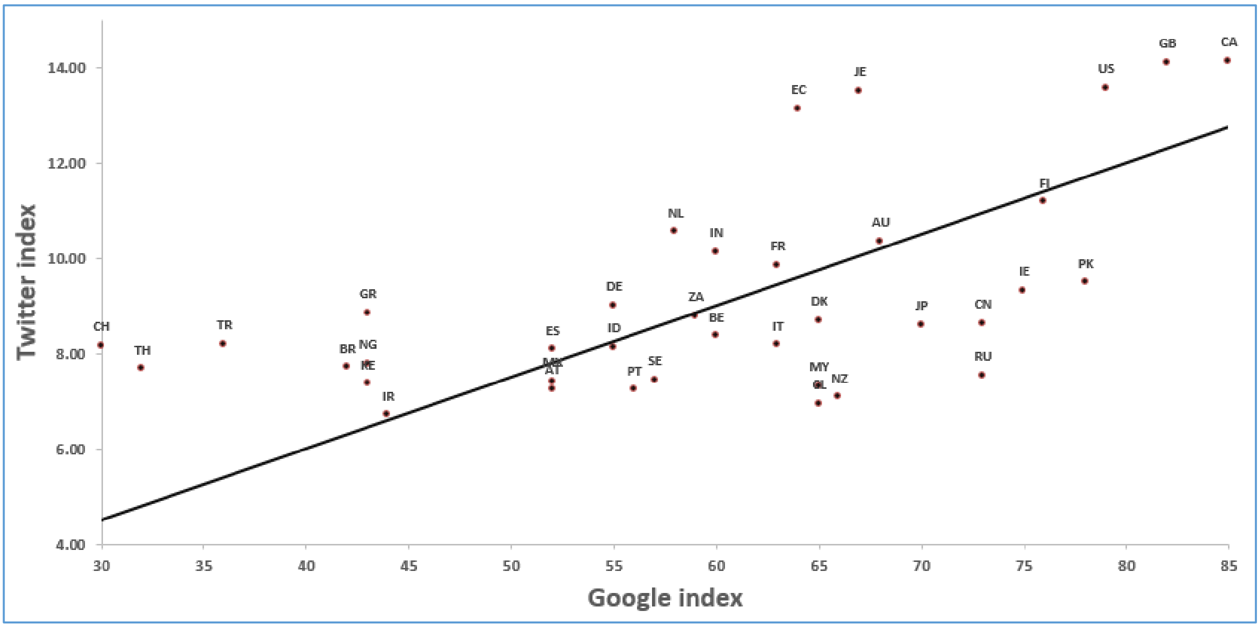


Figure 4. Correlation between Twitter index with Google index.

We then measured the coefficient of correlation between the Google index value and the Twitter index. The result was 0.685, which indicates a strong correlation [61].

9.1.2 Friendliness of Countries

The friendliness of a country is calculated by using Equation 1. It is the average sentiment score for each country. Tables 5 and 6 show information about the most and least friendly countries; respectively, along with tweets statistics. Figure 5 plots the friendliness scores for the top twenty countries. Table 6 lists the least friendly countries.

The top friendly countries were Finland, Brazil and Thailand. The least friendly countries were Switzerland, Austria and Russia. Of the top twenty countries, Figure 5 shows that only five countries have friendliness scores over zero, while the rest have below-zero scores. This result indicates that the public opinion is still highly negative towards Palestine even in the top friendly countries. Several countries like France, Greece, Nigeria and Italy got close to zero friendliness scores.

Referring to the distribution of sentiments and the standard deviation in Tables 5 and 6, a high divergence of attitudes can be observed in most countries. For countries like France, Italy and the UK, the numbers of positive and negative tweets were mostly comparable, while neutral tweets were much fewer in numbers. This result shows that the public opinion in these countries is highly divided. The small number of neutral voices also indicates the large polarization in the public opinion towards the Palestinian issue.

9.1.3 Time-based Analysis

The motivation of time-based analysis is to explore how the public opinion varies over time. Each tweet in our dataset is associated with a timestamp that specifies when the tweet was posted. Therefore, tweets can be treated as time series that can be analysed to extract meaningful patterns.

Due to the variations among countries, utilizing the whole volume of tweets for time-based analysis can result in a large variance. Therefore, time-based analysis was carried out only for the top three countries in terms of the number of posted tweets. These countries are Canada, the UK and the US.

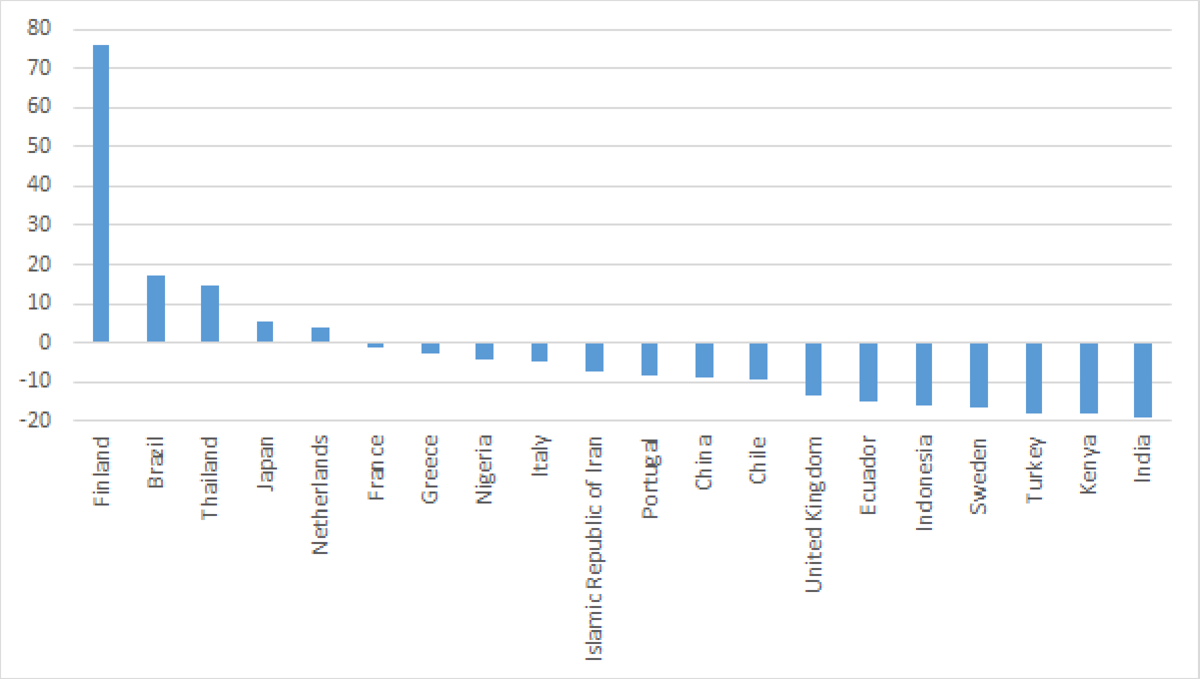


Figure 5. Friendliness scores of top twenty friendly countries.

Table 5. Top ten countries in terms of friendliness.

No.	Country	Focused Tweets	Positive	Negative	Neutral	**Friendliness**	St. Dev.
1	Finland	3,654	3,177	401	76	**75.97**	0.63
2	Brazil	382	184	118	80	**17.28**	0.87
3	Thailand	262	127	89	46	**14.50**	0.90
4	Japan	642	308	272	62	**5.61**	0.95
5	Netherlands	2,646	1,182	1,081	383	**3.82**	0.92
6	France	1,215	440	457	318	**-1.40**	0.86
7	Greece	820	317	338	165	**-2.56**	0.89
8	Nigeria	315	104	118	93	**-4.44**	0.84
9	Italy	577	207	235	135	**-4.85**	0.87
10	Islamic Republic of Iran	218	80	96	42	**-7.34**	0.90

Figure 6 shows how the friendliness scores of these countries have changed over the year 2016. It is obvious that the public opinion in the three countries fluctuated over time and the pattern of change was similar for the three countries. Friendliness scores were low in the first half of the year, before rising up to a peak value in June-July. Attitudes then went down again, then went up at the end of September, before going down again.

To understand these results, we tried to link the time-based changes with the significant events that took place over the year and that were related to the Palestinian-Israeli issue. These events could be discovered easily by searching the news archives on the Web. The declining attitude in the first quarter of 2016 may be explained by the stabbing spree that took place in Jerusalem and other Palestinian cities. A total number of 354 tweets related to the stabbing incidents were tweeted in the three countries during

Table 6. Bottom ten countries in terms of friendliness.

No.	Country	Focused Tweets	Positive	Negative	Neutral	**Friendliness**	St. Dev.
1	Switzerland	381	72	258	51	**-48.82**	0.79
2	Australia	3,125	754	1,862	508	**-35.46**	0.84
3	United States	20,125	4,762	10,203	5,160	**-27.75**	0.82
4	South Africa	717	171	370	176	**-27.04**	0.82
5	Russian Federation	257	60	125	72	**-25.29**	0.81
6	New Zealand	177	44	88	45	**-24.86**	0.83
7	Belgium	399	90	186	123	**-24.06**	0.80
8	Mexico	278	65	131	82	**-23.74**	0.81
9	Germany	830	225	416	189	**-23.01**	0.85
10	Denmark	639	177	322	140	**-22.69**	0.85

the first quarter of 2016. The rising attitudes towards Palestine in June-July 2016 may be attributed to the demolitions of Palestinian houses that took place in July 2016 and resulted in the displacement of dozens of Palestinians(2). In addition, the press releases that accused Israel of forcing Palestinians to withstand cruel and inhuman conditions at its borders have also grabbed attention during June 2016(3)(4). In total, 388 tweets were posted in response to the former events in June-July 2016.

Another rise of attitude towards Palestine was observed in September 2016 that can be explained by the reaction over the death of Shimon Peres, the former Israeli Prime Minister who is seen as a war criminal by pro-Palestinians(5). In total, 571 tweets referring to "Shimon Perez" were tweeted from these countries during September-October 2016, most of which had positive polarity with respect to Palestine. In addition, the UNESCO resolution on 12th October 2016 that condemned the Israeli policies around Al-Aqsa Mosque compound also got considerable attention in social media(6)(7). 332 tweets related to the UNESCO resolution were tweeted from the three countries during October-November 2016.

9.2 Individual-level Analysis

Individual-level analysis aims to explore the attitudes of specific types of individuals. Two groups of individuals will be identified: opinion leaders and individuals with certain ethnicities.

9.2.1 Influence of Opinion Leaders

Different metrics haven been used in the literature to identify opinion leaders on social networks. Some of these metrics have utilized the number of followers, interactions and activity, the leadership or social network analysis [44, 62, 63]. In this work, opinion leaders will be identified by using the number of

(2) http://www.aljazeera.com/news/2016/07/israeli-demolitions-displace-dozens-palestinians 160713124336539.html

(3) http://www.independent.co.uk/news/world/middle-east/israel-border-crossing-checkpoint-palestinians-west-bank-btselem-a7106486.html

(4) http://www.btselem.org/press_releases/20160727_house_demolitions_in_area_c

(5) http://www.independent.co.uk/news/world/middle-east/jerusalem-mayor-palestinians-animals-terror-attack-two-killed-meir-turgeman-a7355116.html

(6) https://www.middleeastmonitor.com/20161013-unesco-vote-no-link-between-al-aqsa-and-judaism/

(7) http://www.aljazeera.com/news/2016/10/palestinians-unesco-vote-al-aqsa-compound-161015133808135.html

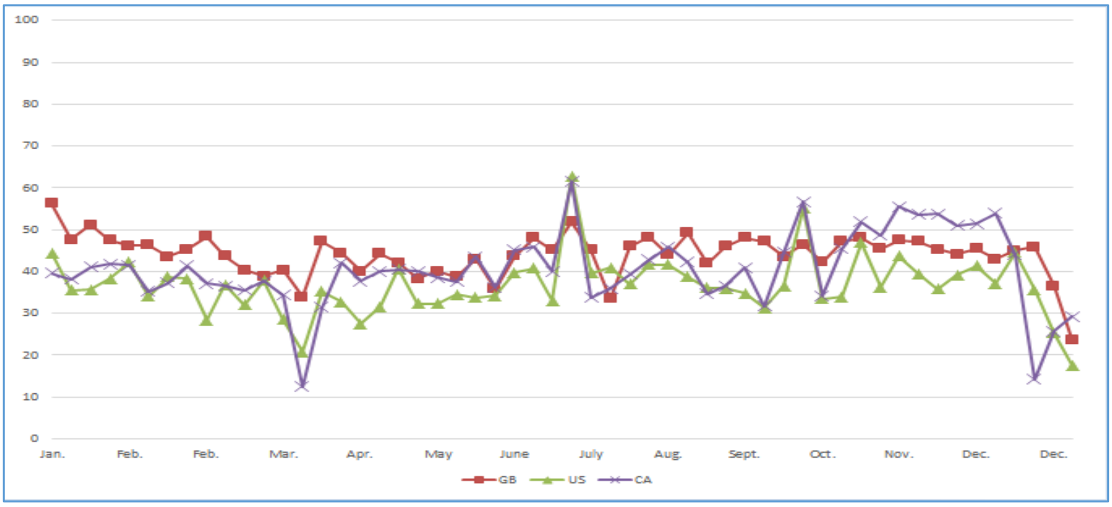


Figure 6. Time-based analysis of public opinion in three countries (UK, US and Canada).

followers, so that users with the largest number of followers in each country will be treated as opinion leaders.

We used the method proposed by Moore and McCabe [64] to identify users with extreme number of followers in each country. Moore and McCabe's method has been widely used in data analysis to find outliers in a distribution, whereas an outlier is the number that is more than 1.5 times the length of the box away from either the lower or upper quartiles. In our approach, opinion leaders are Twitter users whose numbers of followers are considered as "outliers above the upper quartiles" based on the Moore and McCabe's method.

From a total of 38,328 users, 1,794 users were identified as opinion leaders. Table 7 shows statistics about the opinion leaders, while Table 8 shows the top ten countries in terms of the number of opinion leaders. The US, Canada and the UK have the majority of opinion leaders; i.e., 59.14%.

Table 7. Statistics of opinion leaders.

Avg. no. of followers per opinion leader	203623.49
St. dev. of followers per opinion leader	89015.57
Avg. no. of tweets per opinion leader	13.76

Table 8. Top 10 countries in terms of number of opinion leaders.

No.	Country	No. of opinion leaders
1	United States	425
2	Canada	391
3	United Kingdom	299
4	France	95
5	India	61
6	Pakistan	35
7	Finland	31
8	Australia	29
9	Netherlands	23
10	South Africa	22

Identified leaders were mostly official organizations, such as newspapers, government officials or media personnel. For example, among the top opinion leaders in the US were Reuters, Bernie Sanders and Billboard, while among the top opinion leaders in the UK were The Economist, ABC News and United Nations.

After identifying opinion leaders, the friendliness scores for them were calculated by using Equation 1. Then, the average friendliness score of opinion leaders in each country was calculated. The standard deviation per country was also calculated to identify the variance in friendliness of opinion leaders.

Figure 7 shows the results for the top twenty countries in terms of friendliness of opinion leaders, while Figure 8 shows the standard deviation values for friendliness of opinion leaders.

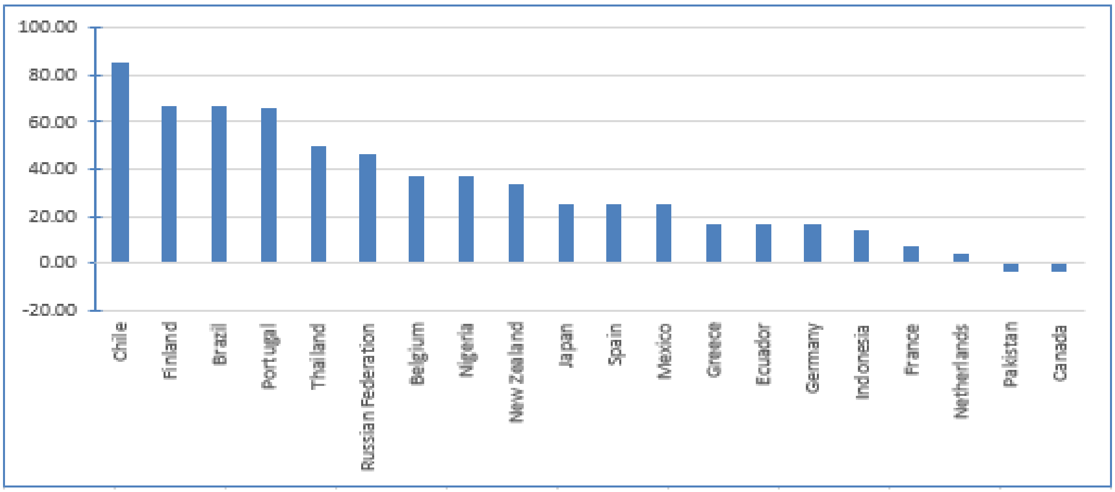


Figure 7. Average friendliness scores of opinion leaders per country.

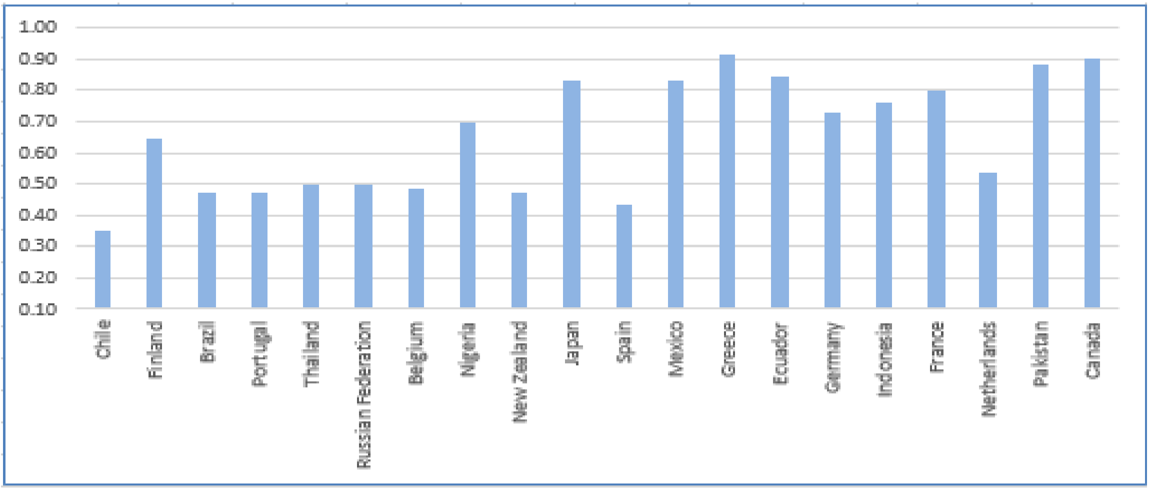


Figure 8. Standard deviation for friendliness of opinion leaders.

Results show that opinion leaders from Chile, Finland and Brazil had the most favourable views of Palestine. It is also noticed that countries that posted the most tweets; i.e., Canada, the US and the UK, are ranked low in terms of the friendliness of their opinion leaders. Looking at the standard deviations, the variation among opinion leaders increases when the friendliness score is low and *vice versa*. This indicates that opinion leaders were highly divided over the Palestinian-Israeli conflict. For example, the variance is high in countries like Germany and Canada in which the friendliness scores are low, while the variance is low in Chile and Spain.

Figure 9 shows the friendliness of opinion leaders as compared to the friendliness of the top twenty countries that generated the most tweets. In general, the attitude of opinion leaders looks consistent with the attitude of their countries for most countries. However, opinion leaders have a slightly more positive attitude towards Palestine as compared to the attitude of the public opinion as in the cases of the UK, Brazil and Chile. On the contrary, countries like Japan, France and China have leaders with less favourable views towards Palestine as compared to the country's friendliness score.

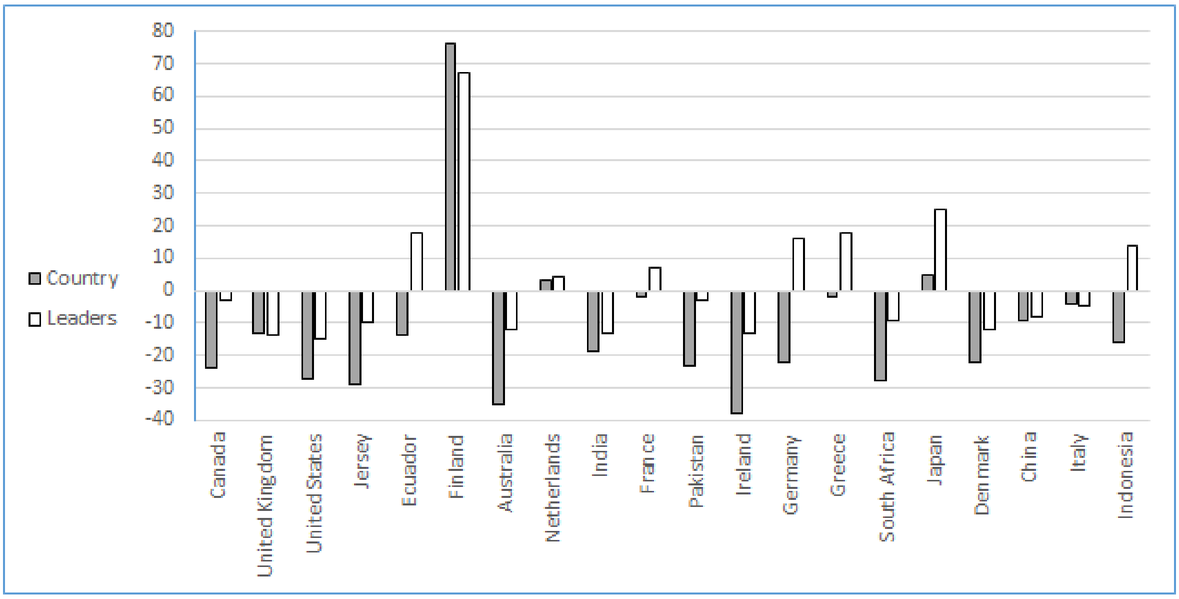


Figure 9. Friendliness of countries *vs.* opinion leaders.

9.2.2 Influence of Individual's Background

Individuals who share the same ethnicity, race or religion are likely to be sympathetic to each other’s issues. For example, a large number of Muslim and Arab people living in Europe and North America provide continuous support to the Palestinian people. Part of this support comes through social networks in different forms, such as retweet campaigns, hashtags, fundraising and promoted tweets. The positive attitude of Muslim and Arab individuals is largely driven by shared culture or religious motivation.

The friendliness scores presented in Table 5 show the overall country's attitude, but do not show how this attitude is influenced by the background of Twitter users or how different groups, such as Muslims or Arabs, contribute to the public opinion in their countries. Identifying the attitudes of different groups will be helpful for decision makers and social media activists, so that they can alter their speech and dialogue according to the needs and motivates of each group.

For simplicity, we decided to classify Twitter users from each country into two groups based on their names: a group of people who have Arabic or Muslim names (we refer to it as "Arab_Muslims" group) and a group of people who have other names (we refer to it as "non-Arab_Muslim" group). One should note here that the group with Arabic names is not restricted to Arab people, but may include people from the wider Muslim world, such as Pakistanis, Iranians and Indians. Usernames can give an indicator of the ethnic or religious group to which a Twitter user belongs. However, the limitation of using usernames is that some Twitter users may use nicknames not related to their original names.

To identify Arabic names, we used a dictionary of Arabic names that we constructed from[8]. In total, 828 Arabic names were included in the dictionary, besides the different ways of writing these names in both Arabic and English. Each Twitter username in the collected tweets was compared with the names in the dictionary. If the username contains an Arabic name, in either Arabic or English, it is added to the group of people with Arabic names. Otherwise, it is assumed to be a non-Arab user. After identifying users with Arabic and non-Arabic names, friendliness scores of each group is calculated.

Table 9 presents the analysis results for the top twenty countries that posted the most tweets. For each country, the friendliness score is presented along with the friendliness of the "Arab_Muslim" and "non_Arab_Muslim" groups. The percentage of tweets posted by each group is also shown. In general, the contribution of the Arab-Muslim communities was marginal for most countries as is evident from

[8] Behind the Name: the etymology and history of first names, https://www.behindthename.com (Last Access: 15/04/2019)

the small numbers of tweets posted by them. This result is expected, because Muslims and Arabs are minorities in most surveyed countries.

Users with Arabic names have more favourable views of Palestine. Friendliness was high among Arabs and Muslims in most major Western countries, such as Canada, the UK and the US. The friendliness scores were low or even negative in countries like Japan and South Africa. This result does not necessarily reflect the situation, because the number of tweets identified as being posted by Arabs and Muslims in these countries was too small to be representative of the entire population.

In general, the positive sentiment of Arabs and Muslims in most countries did not influence the overall public opinion due the small number of tweets. Apart from Finland, users with non-Arabic names have negative friendliness scores.

Table 9. Friendliness of user groups for the countries that posted the most tweets.

Country Name	Country's Friendliness	"Arab_Muslim" group		"Non_Arab_Muslim" group	
		Friendliness	Percentage of tweets	**Friendliness**	Percentage of tweets
Canada	-24.43	**50.76**	2.8%	**-26.65**	97.2%
United Kingdom	-13.31	**27.51**	1.7%	**-14.01**	98.3%
United States	-27.04	**34.51**	0.2%	**-28.64**	99.8%
Jersey	-29.16	**58.33**	0.1%	**-29.25**	99.9%
Ecuador	-14.88	**7.14**	0.45%	**-14.87**	99.55%
Finland	75.97	**20.00**	0.1%	**76.05**	99.9%
Australia	-35.46	**22.64**	1.7%	**-36.46**	98.3%
Netherlands	3.82	**7.14**	1%	**3.78**	99%
India	-19.38	**1.10**	6.3%	**-20.75**	93.7%
France	-1.40	**23.08**	2.1%	**-1.93**	97.9%
Pakistan	-23.89	**26.79**	21.52%	**-37.80**	78.48%
Ireland	-37.38	**12.50**	0.82%	**-37.80**	99.18%
Germany	-23.01	**3.33**	3.61%	**-24.00**	96.39%
Greece	-2.56	**9.52**	2.56%	**-2.88**	97.44%
South Africa	-27.75	**-2.70**	5.16%	**-29.12**	94.84%
Japan	5.61	**0.00**	0.31%	**5.63**	99.69%
Denmark	-22.69	**4.55**	3.44%	**-23.66**	96.56%
China	-9.12	**3.09**	15.25%	**-11.32**	84.75%
Italy	-4.85	**20.00**	2.6%	**-5.52**	97.4%
Indonesia	-16.31	**14.29**	2.7%	**-17.16**	97.3%

10. Conclusions and Future Work

This research proposes an approach for political sentiment analysis at both country and individual levels. The approach was implemented to explore the international public opinion towards the Palestinian-

Israeli conflict by using Twitter data. A dataset consisting of 178,524 tweets posted during 2016, was collected and pre-processed. The polarities of tweets were first measured by using a sentiment analyser that was specially trained to identify the sentiment about Palestine. Several features were then extracted and analyzed to provide a deep insight into the public opinion.

There are many directions to extend this work in the future: First, we aim to use a larger dataset of tweets that span over several years. This will likely generate more reliable and generalizable results. Second, we aim to improve the sentiment analyser by training it with a larger volume of tweets. This is crucial, because the whole analysis is based on the polarities generated by the sentiment analyser. Third, we aim to explore and use more reliable approaches to identify opinion leaders and individual's background and characteristics. Forth, we plan to perform content analysis by means of topic modelling in order to explore what people are discussing with respect to the Palestinian-Israeli conflict.

We think that other researchers, not necessarily from the IT discipline, can also build on these results to gain deeper insights. For example, the results from this analysis may be compared with the results of related national polls in order to explore similarities and/or differences.

REFERENCES

[1] A. Bermingham and A. Smeaton, "On Using Twitter to Monitor Political Sentiment and Predict Election Results," Proceedings of the Workshop on Sentiment Analysis Where AI Meets Psychology (SAAIP 2011), pp. 2-10, 2011.

[2] A. Tumasjan, T. O. Sprenger, P. G.Sandner and I. M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, Association for the Advancement of Artificial Intelligence, pp. 178-185, 2010.

[3] A. Jungherr, "Twitter Use in Election Campaigns: A Systematic Literature Review," Journal of Information Technology & Politics, vol. 13, no. 1, pp. 72-91, 2016.

[4] B. O'Connor, R. Balasubramanyan, B. R. Routledge and N. A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, Association for the Advancement of Artificial Intelligence, pp. 122-129, 2010.

[5] H. Wang, D. Can, A. Kazemzadeh, F. Bar and S. Narayanan, "A System for Real-time Twitter Sentiment Analysis of 2012 US Presidential Election Cycle," Association for Computational Linguistics, pp. 115-120, 2012.

[6] E. Siapera, "Tweeting# Palestine: Twitter and the Mediation of Palestine," International Journal of Cultural Studies, vol. 17, no. 6, pp. 539-555, 2014.

[7] E. Siapera, G. Hunt and T. Lynn, "# GazaUnderAttack: Twitter, Palestine and Diffused War," Information, Communication & Society, vol. 18, no. 11, pp. 1297-1319, 2015.

[8] P. Sobkowicz, M. Kaschesky and G. Bouchard, "Opinion Mining in Social Media: Modeling, Simulating and Forecasting Political Opinions in the Web," Government Information Quarterly, vol. 29, no. 4, pp. 470-479, 2012.

[9] J. A. Balazs and J. D. Velásquez, "Opinion Mining and Information Fusion: A Survey," Information Fusion, vol. 27, pp. 95-110, 2016.

[10] E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Urena-López and A. R. Montejo-Ráez, "Sentiment Analysis in Twitter," Natural Language Engineering, vol. 20, no. 1, pp. 1-28, 2014.

[11] F. Bellini and N. Fiore, "Exploring Sentiment on Financial Market Through Social Media Stream Analysis," Reshaping Accounting and Management Control Systems, Springer, pp. 115-129, 2017.

[12] M. Nardo, M. Petracco-Giudici and M. Naltsidis, "Walking Down Wall Street with a Tablet: A Survey of Stock Market Predictions Using the Web," Journal of Economic Surveys, vol. 30, no. 2, pp. 356-369, 2016.

[13] J. Bollen, H. Mao and X. Zeng, "Twitter Mood Predicts the Stock Market," Journal of Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[14] P. N. Howard, G. Bolsover, B. Kollanyi, S. Bradshaw and L.-M. Neudert, "Junk News and Bots during

the US Election: What Were Michigan Voters Sharing over Twitter," Computational Propaganda Research Project, Oxford Internet Institute, Data Memo, 2017.

[15] D. Murthy, "Twitter and Elections: Are Tweets, Predictive, Reactive or a Form of Buzz?," Information, Communication & Society, vol. 18, no. 7, pp. 816-831, 2015.

[16] P. L. Francia, "Free Media and Twitter in the 2016 Presidential Election: The Unconventional Campaign of Donald Trump," Social Science Computer Review, vol. 36, no. 4, pp. 440-455, 2018.

[17] D. Gayo-Avello, "No, You Cannot Predict Elections with Twitter," IEEE Internet Computing, vol. 16, no. 6, pp. 91-94, 2012.

[18] P. Howard, B. Kollanyi and S. C. Woolley, "Bots and Automation over Twitter during the Second US Presidential Debate," The Computational Propaganda Project, Oxford Internet Institute, pp. 1-4, 2016.

[19] D. Freelon and D. Karpf, "Of Big Birds and Bayonets: Hybrid Twitter Interactivity in the 2012 Presidential Debates," Information, Communication & Society, vol. 18, no. 4, pp. 390-406, 2015.

[20] K. Gorkovenko and N. Taylor, "Understanding How People Use Twitter during Election Debates," Proceedings of the ACM 31st British Computer Society Human Computer Interaction Conference, BCS Learning & Development, Ltd., pp. 88, 2017.

[21] D. Kreiss, "Seizing the Moment: The Presidential Campaigns' Use of Twitter during the 2012 Electoral Cycle," New Media & Society, vol. 18, no. 8, pp. 1473-1490, 2016.

[22] T. Lansdall-Welfare, F. Dzogang and N. Cristianini, "Change-point Analysis of the Public Mood in UK Twitter during the Brexit Referendum," Proc. of IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 434-439, 2016.

[23] N. A. Diakopoulos and D. A. Shamma, "Characterizing Debate Performance *via* Aggregated Twitter Sentiment," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp. 1195-1198, 2010.

[24] S. Stieglitz and L. Dang-Xuan, "Political Communication and Influence through Microblogging: An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior," Proc. of the 45th IEEE Hawaii International Conference on System Sciences, pp. 3500-3509, 2012.

[25] I. Alfina, D. Sigmawaty, F. Nurhidayati and A. N. Hidayanto, "Utilizing Hashtags for Sentiment Analysis of Tweets in the Political Domain," Proceedings of the 9th International Conference on Machine Learning and Computing, ACM, pp. 43-47, 2017.

[26] P. Grover, A. K. Kar, Y. K. Dwivedi and M. Janssen, "Polarization and Acculturation in US Election 2016 Outcomes–Can Twitter Analytics Predict Changes in Voting Preferences," Technological Forecasting and Social Change, pp. 1-23, 2018.

[27] R. Bose, R. K. Dey, S. Roy and D. Sarddar, "Analyzing Political Sentiment Using Twitter Data," Information and Communication Technology for Intelligent Systems, Springer, pp. 427-436, 2019.

[28] D. Paul, F. Li, M. K. Teja, X. Yu and R. Frost, "Compass: Spatio Temporal Sentiment Analysis of US Eelection What Twitter Says!," Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 1585-1594, 2017**.**

[29] O. Almatrafi, S. Parack and B. Chavan, "Application of Location-based Sentiment Analysis Using Twitter for Identifying Trends towards Indian General Elections 2014," Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication, ACM Edn., 2015.

[30] S. Vosoughi, H. Zhou and D. Roy, "Enhanced Twitter Sentiment Classification Using Contextual Information," arXiv preprint arXiv:1605.05195, 2016.

[31] J. Deegan, J. Hogan, S. Feeney and B. K. O'Rourke, "The Self and Other: Portraying Israeli and Palestinian Identities on Twitter," Irish Communication Review, vol. 16, no. 1, Article 8, 2018.

[32] J. W. Pennebaker, M. E. Francis and R. J. Booth, "Linguistic Inquiry and Word Count: LIWC 2001," Mahway: Lawrence Erlbaum Associates, vol. 71, 2001.

[33] P. Burnap, R. Gibson, L. Sloan, R. Southern and M. Williams, "140 Characters to Victory?: Using Twitter to Predict the UK 2015 General Election," Electoral Studies, vol. 41, pp. 230-233, 2016.

[34] A. Ceron, L. Curini and S. M. Iacus, "Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters—Evidence from the US and Italy," Soci. Sci. Comp. Review, vol. 33, no. 1, pp. 3-20, 2015.

[35] A. Ceron, L. Curini and S. M. Iacus, "To What Extent Sentiment Analysis of Twitter Is Able to Forecast Electoral Results? Evidence from France, Italy and the United States," Proc. of the 7th ECPR General Conference Sciences Po, Bordeaux, pp. 5-8, 2013.

[36] D. J. Hopkins and G. King, "A Method of Automated Non-parametric Content Analysis for Social Science," American Journal of Political Science, vol. 54, no. 1, pp. 229-247, 2010.

[37] H. T. Le, G. Boynton, Y. Mejova, Z. Shafiq and P. Srinivasan, "Revisiting the American Voter on Twitter," Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, ACM, pp. 4507-4519, 2017.

[38] F. Marozzo and A. Bessi, "Analyzing Polarization of Social Media Users and News Sites during Political Campaigns," Social Network Analysis and Mining, Springer, vol. 8, no. 1, 2018.

[39] S. Martin-Gutierrez, J. C. Losada and R. M. Benito, "Recurrent Patterns of User Behavior in Different Electoral Campaigns: A Twitter Analysis of the Spanish General Elections of 2015 and 2016," Complexity Jour., vol. 2018, 2018.

[40] E. M. Cody, A. J. Reagan, L. Mitchell, P. S. Dodds and C. M. Danforth, "Climate Change Sentiment on Twitter: An Unsolicited Public Opinion Poll," PloS One, vol. 10, no. 8, pp. e0136092, 2015.

[41] N. S. Khan, M. Ata and Q. Rajput, "Identification of Opinion Leaders in Social Network," Proc. of IEEE International Conference on Information and Communication Technologies (ICICT), pp. 1-6, 2015.

[42] F. Bodendorf and C. Kaiser, "Detecting Opinion Leaders and Trends in Online Social Networks," Proceedings of the 2nd ACM Workshop on Social Web Search and Mining, pp. 65-68, 2009.

[43] M. Solomon, R. Russell-Bennett and J. Previte, Consumer Behaviour, Pearson Higher Education AU, 3rd Edition , 2012.

[44] F. Riquelme and P. González-Cantergiani, "Measuring User Influence on Twitter: A Survey," Information Processing & Management, vol. 52, no. 5, pp. 949-975, 2016.

[45] M. Gaurav, A. Srivastava, A. Kumar and S. Miller, "Leveraging Candidate Popularity on Twitter to Predict Election Outcome," Proceedings of the 7th Workshop on Social Network Mining and Analysis, ACM, Article 7, 2013.

[46] DNOiSE., "Followthehashtag // Free twitter search analytics and business intelligence tool," [Online], Available: http://www.followthehashtag.com/.

[47] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider and N. A. Smith, "Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters," Proceedings of NAACL-HLT, Association for Computational Linguistics, pp. 380–390, 2013.

[48] K. Gimpel, N. Schneider and B. O'Connor, "Annotation Guidelines for Twitter Part-of-Speech Tagging Version 0.3," [Online], Available: http://www.ark.cs.cmu.edu/TweetNLP/annot_guidelines.pdf.", 2013.

[49] M. J. Idzelis, "The Java Open Source Spell Checker," [Online], Available: http://jazzy.sourceforge.net/, [Accessed: 15-08-2019].

[50] Alias-i., "LingPipe 4.1.0.," [Online], Available: http://alias-i.com/lingpipe, [Accessed: 15-08-2019].

[51] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," pp. 55-60, DOI:10.3115/v1/p14-5010, 2014.

[52] M. Thelwall, K. Buckley, G. Paltoglou, C. Cai and A. Kappas, "SentiStrength," [Online], Available: http://sentistrength.wlv.ac.uk/, [Accessed: 15-08-2019].

[53] R. Kumar, H. S. Pannu and A. K. Malhi, "Aspect-based Sentiment Analysis Using Deep Networks and Stochastic Optimization," Neural Computing and Applications, pp. 1-15, 2019.

[54] A. Valdivia, M. V. Luzón, E. Cambria and F. Herrera, "Consensus Vote Models for Detecting and Filtering Neutrality in Sentiment Analysis," Information Fusion, vol. 44, pp. 126-135, 2018.

[55] A. Milani, N. Rajdeep, N. Mangal, R. K. Mudgal and V. Franzoni, "Sentiment Extraction and Classification for the Analysis of Users' Interest in Tweets," International Journal of Web Information Systems, vol. 14, no. 1, pp. 29-40, 2018.

[56] E. Kušen and M. Strembeck, "Politics, Sentiments and Misinformation: An Analysis of the Twitter Discussion on the 2016 Austrian Presidential Elections," Online Social Networks and Media, vol. 5, pp. 37-50, 2018.

[57] M. Bouazizi and T. Ohtsuki, "Sentiment Analysis: From Binary to Multi-class Classification: A Pattern-based Approach for Multi-class Sentiment Analysis in Twitter," IEEE Access, vol. 5, pp. 1-6, 2016.

[58] U. Yaqub, S. Chun, V. Atluri and J. Vaidya, "Sentiment-based Analysis of Tweets during the US Presidential Elections," Proceedings of the 18th Annual International Conference on Digital Government Research, ACM, pp. 1-10, 2017.

[59] Apache.org, "Apache Spark: Unified Analytics Engine for Big Data," [Online], Available: https://spark.apache.org/, [Accessed: 15/08/2019].

[60] L. Bornmann and R. Haunschild, "How to Normalize Twitter Counts? A First Attempt Based on Journals in the Twitter Index," Scientometrics, vol. 107, no. 3, pp. 1405-1422, 2016.

[61] R. Taylor, "Interpretation of the Correlation Coefficient: A Basic Review," Journal of Diagnostic Medical Sonography, vol. 6, no. 1, pp. 35-39, 1990.

[62] E. Dubois and D. Gaffney, "The Multiple Facets of Influence: Identifying Political Influentials and Opinion Leaders on Twitter," American Behavioral Scientist, vol. 58, no. 10, pp. 1260-1277, 2014.

[63] W. W. Xu, Y. Sang, S. Blasiola and H. W. Park, "Predicting Opinion Leaders in Twitter Activism Networks: The Case of the Wisconsin Recall Election," American Behavioral Scientist, vol. 58, no. 10, pp. 1278-1293, 2014.

[64] D. S. Moore, G. P. McCabe and B. A. Craig, Introduction to the Practice of Statistics, 7th Edition, W.H. Freeman & Company, 2012.

ملخص البحث:

في ظل النمو الهائل لمنصات التواصل الاجتماعي، لا ينشر الناس معلومات عامة فحسب، وإنما ينشرون آراءهم السياسية أيضاً. وقد استخدم الكثير من الأبحاث محتوى وسائل التواصل الاجتماعي لتحليل الرأي العام تجاه الأحداث السياسية والتنبؤ به. يقدم هذا العمل دراسة تحليلية لقياس الرأي العام السياسي تجاه النزاع الفلسطيني – الإسرائيلي باستخدام بيانات تويتر. تستخدم هذه الدراسة نموذجاً مبتكراً لتحليل البيانات يُعنى بمستويين من التحليل هما: التحليل على مستوى الدولة، والتحليل على مستوى الفرد. يهدف التحليل على مستوى الدولة الى استكشاف الاتجاه الإجمالي للدّول تجاه فلسطين، وذلك عبر: 1. تحديد الدول التي صدر منها أكثر التغريدات المتعلقة بالموضوع؛ 2. قياس دراجة الصداقة لكل دولة نحو فلسطين؛ 3. تحليل التغير في الرأي مع الوقت. أما التحليل على المستوى الفردي فيهدف الى تحليل البيانات بناءً على نشاط الأفراد وخلفياتهم. وقد جرى تحليل اتجاهات كل من قادة الرأي والمجموعات الإثنيّة ومناقشتها في ضوء اتجاهات الدول.

إن التجربة الغنية التي يقدمها هذا البحث من خلال النموذج المقترح، والإجراءات التي اتُّبعت خطوة بخطوة، وتنوع تقنيات التحليل، ومناقشة النتائج، من شأنها أن تضع معلومات وافرة بين أيدي مطوري النماذج ومحللي البيانات المهتمين في تحليل الآراء التي يعبر عنها على منصات التواصل، فيما يتعلق بالنزاعات السياسية على وجه الخصوص.

BENCHMARKING MACHINE LEARNING ALGORITHMS FOR ANDROID MALWARE DETECTION

Somayyeh Fallah[1] and Amir Jalaly Bidgoly[2]

(Received: 3-Jun.-2019, Revised: 31-Jul.,-2019 and 12-Aug.-2019, Accepted: 4-Sep.-2019)

ABSTRACT

Nowadays, smartphones have captured a significant part of human life and has led to an increasing number of users involved with this technology. The rising number of users has encouraged hackers to generate malicious applications. Identifying these malwares is critical for preserving the security and privacy of users. The recent trend of cyber security shows that threats can be effectively identified using network-based detection techniques and machine learning methods. In this paper, several well-known methods of machine learning were investigated for smartphone malware detection using network traffic. A wide range of malware families are used in the investigations, including Adware, Ransomware, Scareware and SMS Malware. Also, the most used and famous supervised and unsupervised machine learning methods are considered. This article benchmarked the methods from different points of view, such as the required features count, the recorded traffic volume, the ability of malware family identification and the ability of a new malware family detection. The results showed that using these methods with appropriate features and traffic volume would achieve the F1-measure of malware detection by a percentage of about 90%. However, these methods did not show acceptable results in detecting malicious as well as new families of malware. The paper also explained some of the challenges and potential research problems in this context which can be used by researchers interested in this field.

KEYWORDS

Android malware, Malware detection, Network traffic, Machine learning.

1. INTRODUCTION

With the enormous growth of smartphones around the world, the number of malware attacking mobile applications has also witnessed an exponential increase. Due to their wide variety and open source platform, these phones have become an attractive domain for hackers to penetrate. According to a report by F-Secure Corporation in 2017 [1], more than 99 percent of all malware designed for smartphones target Android devices. There are over 19 million malware programs developed particularly for Android, making Google’s mobile operating system the main target for mobile malware. The reason for this is the vast distribution of Android devices, as well as the relatively open system for the distribution of apps. Many malware programs use the Internet in order to communicate with the initiator of the attack in order to receive new tasks and software updates or to leak collected data. Yet, when such malware tries to communicate with its Command and Control (C&C) server, it most likely uses a common and known network protocol to pass through firewalls [2]. By studying, capturing and analyzing the flow of information between two hosts, network administrators are able to provide a basic behavioral pattern. When they are familiar with network behavior, they can catch anomalies, such as significant increases in bandwidth usage, distribution of DDOS attacks and other unauthorized occurrences. By analyzing network traffic and identifying suspicious domains, network administrators can detect malware infections months before the actual malware would be discovered. This could be indicative of the fact that malicious attackers need to communicate with their command and control unit of computers, creating network traffic that can be identified and analyzed. Having a previous warning of developing malware infections can enable faster responses and reduce the impact of attacks. Network traffic classification is the first step in analyzing and identifying different types of programs running on a network. Through this method, Internet service providers or network operators can manage the overall network performance. There are three main approaches to traffic classification: 1) port-based approaches, 2) payload-based approaches and 3) machine learning approaches [3]-[4]. The most common and promising approach in the field of traffic classification is the use of machine learning methods. These methods, which are also able to overcome the constraints of both port-based and

S. Fallah and A. Jalaly are with Department of Computer Engineering, University of Qom, Iran. Emails: [1]s.fallah@stu.qom.ac.ir and [2]jalaly@qom.ac.ir

payload-based methods, assumed that applications send data with a regular pattern. These patterns can be used as a means to identify traffic categories. To find these patterns, flow statistics (such as the average packet size, flow lengths and total number of packets) and just the use of the TCP protocol can be effective in the classification process [5].

Machine learning is a data analysis method that automatically performs analytical modeling. This method is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Factors, such as increasing growth in data types, powerful and inexpensive computing process and the storage of cost-effective data have led to rapid and automatic development of models that are capable of analyzing large and complex data, delivering faster and more accurate results and causing a wave of interest and popularity.

In this paper, with the aim of identifying Android malware, we examine the impact of machine learning methods on more accurate detection of the presence of malware in smartphones. For this reason, we present a comprehensive evaluation of all aspects that are effective in achieving an acceptable result. In this evaluation, it is attempted to take a deep look into various aspects of machine learning methods in order to present their strengths and weaknesses. These evaluations are important in providing researchers with useful and comprehensive information for conducting research work in this field. It matters to other researchers in large scale as well. These evaluations include:

1.1 Benchmarking in Terms of the Number of Extracted Features and the Selection of High-grade Features

The purpose of this evaluation is to provide a subset of features that, in addition to feature reduction and optimality of the problem, can identify normal samples from malicious samples with a very high performance.

1.2 Benchmarking in Terms of Time Duration of Capturing Traffic

Considering the fact that the shorter time duration of detecting malware samples, the lesser impact of infections. This evaluation comparatively analyzes the ability of machine learning methods to detect malware in a shorter time duration from recorded traffic. A method that can detect infection more accurately in the shortest possible time can actually be considered more practical.

1.3 Benchmarking in Terms of the Ability to Identify Malware Families

It is very important that when a malicious program infects the network and the first signals of its presence are observed, the type of malware family should be identified. This makes it possible to prevent any damage by studying the behavior patterns of each family for preventive measures.

1.4 Benchmarking in Terms of the Ability to Identify Unknown Malware Types

In this evaluation, the performance of machine learning methods is examined in the detection of unknown malware types about which no training sample has been given to the system.

In evaluations, all of the features studied in recent research works have been used. This set of selective features provided acceptable results in previous studies. Hence, after the creation of dataset, a model is designed which can provide the best prediction for identifying malicious traffic patterns from normal traffic patterns. The results of algorithms act as a criterion to measure the performance of machine learning algorithms.

The paper continues as follows: Section 2 reviews the related works on malware detection using network traffic. Section 3 describes the general explanation of how to perform the benchmarks. The benchmarks are discussed in section 4. Section 5 discusses the limitations and threats to the validity of the results and finally, the paper ends in Section 6 with the research conclusions.

2. LITERATURE REVIEW

The malware detection method in Android platform is in two broad categories: static analysis and dynamic analysis [6]. In static analysis, no application is executed; only the code and other components, like manifest files, are analyzed. Therefore, it is a quick and inexpensive approach, whereas in dynamic analysis, the applications are executed on actual or virtual environments and normal programs are distinguished from malicious programs using event logs, memory, processor and network usage and analysis. The use of dynamic methods to examine network behavior and its evaluation by machine learning methods has become a popular approach to identify and categorize the malware family. A lot of related work has been focused on this topic, which was discussed in [7]-[9]. In a work by Arora in 2014 [10], the network traffic features are analyzed and a rule-based classification is created to detect Android malware. Then, it provides a list of features that can distinguish between malware traffic and normal traffic patterns and a rule-based classification is built on the acquired features. The results of this study indicated that this approach is significantly correct and identifies more than 90% of the traffic samples. In another work [11], an unsupervised machine learning approach is used for Internet traffic identification and the results are compared with a supervised machine learning approach. The unsupervised approach uses a clustering algorithm and the supervised approach uses the Naive Bayes classifier. Finally, it is concluded that the unsupervised clustering technique has an accuracy up to 91% and performed up to 9% better than the supervised technique. The authors of the previous paper, in another study [12], evaluated network traffic with three unsupervised algorithms: K-Means, DBSCAN and Auto Class. The experimental results showed that the Auto Class algorithm produces the best overall accuracy and with a very small difference, the accuracy of the K-Means algorithm is less than that of the Auto Class algorithm, but, similar to the DBSCAN algorithm, it has a high speed in designing the model. In another work [14], to overcome the drawbacks of existing methods for traffic classification, usage of C5.0 Machine Learning Algorithm (MLA) was proposed. Based on traffic statistics, an advanced classifier was constructed which was able to distinguish between 7 different applications in the test set of 76,632–1,622,710 unknown cases with an average accuracy of 99.3%–99.9%. In a paper by Alhawi [15], a machine learning evaluation study for consistent detection of windows ransomware network traffic was introduced. Using a dataset created from conversation-based network traffic features, a True Positive Rate (TPR) of 97.1% was achieved by the Decision Tree (J48) classifier. The training set included 75618 samples and a test set consisting of 45526 samples. In this experiment, six classifiers of the *Bayesian* network, random forest, KNN, J48, multilayer perceptron and logistic model tree (LMT) were used.

In an article written by Pendlebury et al. [16], it is shown that the results of malware classification can be affected for two reasons: spatial bias caused by distributions of training and testing data that are not representative of a real-world deployment and temporal bias caused by incorrect time splits of training and testing sets, leading to impossible configurations. Hence, a set of space and time constraints for an experiment design that eliminates both sources of bias was proposed. A new metric that summarizes the expected robustness of a classifier in a real-world setting was introduced and provided an algorithm for its performance. An open source evaluation framework called TESSERACT was used to compare the three malware classifiers (decision tree, SVM and deep learning), finally shown that TESSERACT is fundamental to accurate evaluation and comparison of different solutions, especially when considering mitigation strategies for time decay. The dataset used consisted of 129K applications. In another work [17], the network traffic was used as a dynamic feature Android malware detection. A set including 16 features which can distinguish between normal and malicious traffic was determined. Decision tree classifier was built on top of these distinguishing features only and a set of 217 samples was given as input to the classifier. The results of this study indicated that this approach identifies more than 90% of the traffic samples. In a work by Chen and his colleagues [18], statistical features of mobile traffic are utilized to identify malicious traffic flows. After analyzing traffic flow statistics, the data imbalance issue is detected that significantly affects the performance of Android malicious flow detection. Based on six network flow features extracted from the flow data set, several classification algorithms; namely, Bayes Net, SVM, C4.5, Grading, Ada boost and Naïve Bayes, were implemented and experiments were performed on various imbalanced datasets with different imbalance ratio (IR) values. The results show that most of the commonly-used classifiers achieve reasonable performance, which confirms that machine learning algorithms are effective in identifying malicious mobile traffic. Then, the performance of the IDGC model is examined in addressing the data imbalance issue. After testing the IDGC model on the same traffic flow dataset, it was shown that the IDGC is significantly more stable than other classifiers. By increasing the IR value, the performance of the IDGC classification is maintained for the AUC and GM range between 0.8 and 1.0. But, the IDGC classification process is very time-consuming, which makes real-time detection impractical. To improve the performance of the IDGC model, the authors proposed a novel S-IDGC model, which optimizes the weight coefficients using an efficient simplex method. The evaluation results showed that S-IDGC inherits the stability characteristic of the IDGC model while drastically reducing time consumption (with approximately 17 times improvement compared with IDGC on time efficiency). Another study [19] demonstrated a behavioral detection method for detecting mobile malware that can communicate with blacklisted domains and pass sensitive personal / financial information. First, an App-URL table is created that logs all attempts made by all applications to communicate with remote servers. Each entry in this log preserves the application id and the URL, so that the application is contacted. From this log, with the help of a reliable and comprehensive domain blacklist, malicious applications that communicate with malicious domains can be detected. In [20], the validation of machine learning malware detection is discussed with in the lab and in the wild scenarios. At first, a feature set for building classifiers that yields high performance measures in lab evaluation scenarios is tested in comparison with state-of-the-art approaches. To this end, several Machine Learning classifiers that rely on a set of features built from applications' CFGs are devised. They used a sizeable dataset of over 50 000 Android applications collected from sources where state-of-the-art approaches have selected the data. Finally, the authors showed that, in the lab, the proposed approach outperforms existing machine learning-based approaches. However, this high performance does not translate in high performance in the wild. The performance gap was observed, F-measures dropping from over 0.9 in the lab to below 0.1 in the wild. In the work by Chen [21], a framework of utilizing model-based semi-supervised (MBSS) classification on the dynamic behavior data for Android malware detection is proposed. In this paper, focus was on detecting malicious behavior at runtime by using dynamic behavior data for analysis. The main advantage of semi-supervised classification is the strong robustness in performance for out-of-sample testing. The model-based semi-supervised classification uses both labeled and unlabeled data to estimate the parameters, since unlabeled data usually carries valuable information on the model fitting procedure. Specifically, MBSS is compared with the popular malware detection classifiers, such as support vector machine (SVM), k-nearest neighbor (KNN) and linear discriminant analysis (LDA). Finally, it is demonstrated that MBSS has a competitive performance for in-sample classification and maintains strong robustness when applied for out-of-sample testing. Under the ideal classification setting, MBSS has a competitive performance with 98% accuracy and very low false positive rate for in-sample classification.

Some of the work relates to the Intrusion Detection System (IDS), which defines and describes the taxonomies of such systems [22]. In another work by Homoliak [23], the taxonomy of intrusion detection methods is presented. Machine learning-based intrusion detection systems as well as network traffic classification ones are employed. Also, the study described datasets utilized for evaluation of Intrusion Detection Systems (IDSs) and finally provided an overview of obfuscation and evasion approaches in ADS and IDS, supplemented by several prevention techniques against obfuscations and evasions.

The system's features denoted as Advanced Security Network Metrics (ASNMs) are designed and defined and there was a performance improvement of detection by ASNM features and a supervised classifier, resulting in two categories of proposed obfuscation techniques. The first category of obfuscation techniques is called tunneling obfuscation and the second category is called non-payload-based obfuscations, which were evaluated and reviewed.

Considering the importance of traffic analysis and the benefits of machine learning methods to early detection of malware, many recent research works have considered these techniques for malware detection. However, as far as we know, none of these works have yet reviewed the comprehensive aspects of machine learning performance and different goals in malware detection. For example, no research work has discussed whether it is possible to identify the type of malware family or not; nor any of them has ever investigated the timing of the apps and the balance of the classes as effective parameters in the final results. Meanwhile, because the data collection and sampling method is different in each work, it has not been possible to compare the results with each other.

3. METHODOLOGY

The main purpose of this work is to evaluate the effective factors in identifying malware in smartphones, especially Android operating systems. To collect and capture traffic, it is necessary to check the flow of information between the points of connection carefully. In other words, these flows are known as a connection when there is a two-way exchange between two nodes. In fact, a flow is defined by a sequence of packets with the same values for five attributes; namely, Source IP, Destination IP, Source Port, Destination Port and Protocol [24]. More useful information is obtained by checking traffic associated with TCP flows. Hence, only TCP-related packets are considered in this work. In this article, in order to have a complete and comprehensive look at the pattern of behavior of malwares, it has been tried to collect a complete collection of all types of malware families and their behavioral differences are carefully evaluated. Malware samples include four different types of families extracted from the CICAndMal2017 dataset and UNB website [25]. Malware samples in the CICAndMal2017 dataset are classified into four categories: Adware, Ransomware, Scareware and SMS Malware. Our samples come from 42 unique malware families. A collection of benign applications from the Google Play Market was collected in the period from 2015 to 2017. These apps were collected based on their popularity (i.e., top free popular and top free new) for each category available on the market. In order to improve the quality of the surveys, four criteria were used: F1-measure (1), Precision (2). Recall (3) and false positive. Precision is known as positive predictive value **(**PPV**)** and recall is known as sensitivity**,** hit rate and true positive rate **(**TPR**).** F1-measure is a balanced mean between precision and recall. *TP* is the number of positive instances classified correctly; *FP* is the number of negative instances misclassified; *FN* is the number of positive instances misclassified; and *TN* is the number of negative instances classified correctly.

$$F1-measure=\frac{2Precision * Recall}{Precision+\ Recall} \tag{1}$$

$$Precision=\frac{TP}{(TP+FP)} \tag{2}$$

$$Recall=\frac{TP}{(TP+FN)} \tag{3}$$

Since the normal distribution of benign and malware apps in the real world is unbalanced, most machine learning algorithms do not work well with the unbalanced dataset. Unbalanced datasets are a special case of classification problem, where the class distribution is not uniform among the classes. Typically, they are composed of two classes: the majority class and the minority class. This type of sets supposes a new challenging problem for data mining, because standard classification algorithms usually consider a balanced training set and this supposes a bias towards the majority class. To solve the problem of imbalance of data, a stratified sampling method was used. Stratified sampling builds random subsets and ensures that the class distribution in the subsets is the same as in the whole dataset.

The dataset contains 600 samples in the benign class and 400 samples in the malware class. After that, stratified 10-fold cross validation was implemented in our experiment and the average recall per class, average precision per class, FP and F1-measure were used for measuring performance.

To collect the feature set, all of the features used in recent works were studied. In many cases, up to 250 features per flow were extracted. These features are among the most important features of articles [6], [13] and [24]. Ideally, it is best to explore all feature combinations to select the one which gives the best result; however, in practice, this is problematic. For example, having n features, the number of experiments required to conduct training/testing on an algorithm is calculated as follows:

$$\sum_{k=0}^{n}\binom{n}{k}=2^n \tag{4}$$

Therefore, the possible combinations are 2^n, where for n more than 40 the problem becomes unmanageable. For this reason, the forward feature selection algorithm was used to select features in five subsets. In order to have the best feature set, common features between this algorithm and the mentioned articles were chosen. These features are based on five characteristics: behavior-based, byte-based, packet-based, time-based and flow-based. The full list of features is presented in Table 1.

4. BENCHMARKING AND COMPARISON RESULTS

In this section, different classification methods have been comparatively analyzed on the collected data in order to evaluate their strengths.

Table 1. List of network features.

Feature	Description
Behavior-based	
F1	The duration of the flow
Byte-based	
F2	Average number of bytes sent
F3	Average number of bytes received
F4	The total number of bytes used for headers sent
F5	The total number of bytes used for headers received
F6	Ratio of number of incoming bytes to number of outgoing bytes
F7	Average number of bytes per second
Packet-based	
F8	Total number of packets sent
F9	Total number of packets received
F10	Total length of packets sent
F11	Total length of packets received
F12	Average number of packets per second
F13	Average number of packets sent per second
F14	Average number of packets received per second
F15,F16,F17,F18	Min, Mean, Max and standard deviation of the size of packet
F19,F20,F21,F22	Min, Mean, Max and standard deviation of the size of packet sent
F23,F24,F25,F26	Min, Mean, Max and standard deviation of the size of packet received
F27	Average number of packets sent/bulk
F28	Average number of packets received /bulk
F29	Subflow packets sent
F30	Subflow packets received
F31	Ratio of number of incoming packets to number of outgoing packets
Time-based	
F32,F33,F34,F35	Min, Mean, Max and standard deviation time between two packets sent in the forward direction
F36,F37,F38,F39	Min, Mean, Max and standard deviation time between two packets sent in the backward direction
F40,F41,F42,F43	Min, Mean, Max and standard deviation time when a flow was idle before becoming active
F44,F45,F46,F47	Min, Mean, Max and standard deviation time when a flow was active before becoming idle
Flow-based	
F48,F49,F50, F51	Min, Mean, Max and standard deviation of the length of a flow
F52	Average number of packets per flow
F53	Average number of packets sent per flow
F54	Average number of packets received per flow
F55	Average number of bytes sent per flow
F56	Average number of bytes received per flow
F57	The average number of bytes in a subflow in the forward direction
F58	The average number of bytes in a subflow in the backward direction
F59	Variance of total number of bytes used in the backward direction
F60	Variance of total number of bytes used in the forward direction
F61,F62,F63,F64,F65,F66,F67	Number of packets with FIN,SYN,RST,PSH,ACK,URG,CWE
F68	The total number of bytes sent in the initial window in the forward direction
F69	The total number of bytes sent in the initial window in the backward direction
F70,F71,F72	Ave, Max/Min segment size observed in the forward direction
F73,F74,F75	Ave ,Max/Min segment size observed in the backward direction

4.1 Benchmarking Based on the Number of Extracted Features

Feature selection is a very important component in data science. When data is presented on a very large scale, the training time increases resulting in that most models do not have proper and optimum output. Also, the larger dimension feature space creates a larger number of parameters that need to be estimated. As a result, the number of parameters increases with the possibility of overfitting in the model. Appropriate feature selection on one hand leads to a new set of features that are more compact and have more distinctive properties and on the other hand, unrelated and repetitive features are eliminated, which increases the F1-measure of the evaluation. Depending on the type of feature selection algorithm, there are three general approaches to feature selection which include filters, wrappers and embedded [26]-[28]. The wrapper method is widely used for classification issues; therefore, it was used here in this evaluation for feature selection. Features were examined in five subsets:

The first set consisted of 75 features and the second, the third, the fourth and the last sets contained 45, 25, 15 and 9 features, respectively (Table 2). In choosing these sets, packet-based features and then flow-based features were given special priority. These feature categories were included in all the subsets. In packet-based features [29], all received packets are processed; thus producing low false alarms, which makes this method very time-consuming. Flow-based features have an overall lower amount of data to be processed; therefore, this method is the logical choice to work with in high-speed networks. But it has less input information available to detect attacks and suffers from producing high false alarms. The main advantage of packet-based approach is that all common kinds of known attacks and intrusions can be detected if the data source delivers the entire network packet for analysis. On the other hand, performance issues in flow-based method are not a primary concern and therefore it is the logical choice for high-speed networks. Thus, the combination of both features is the best option, as it can detect a wide range of attacks with lowest error and highest speed.

To select the optimal feature set, three feature selection approaches from the wrapper method were investigated. These approaches were: forward selection, backward elimination and optimized selection. For this purpose, in the Rapid Miner tool, these three operators were compared. The survey was carried out each time with several classifications. The results of the survey showed that the optimized selection operator performed the best in all classifiers. Therefore, each feature set was selected by optimized selection.

Since the dataset contains a variety of heterogeneous features, these features are used in various algorithms that require a measure of distance/ similarity. Therefore, the data is normalized before it is used. Normalization is important when dealing with features and data with different scales.

In this evaluation, it has been attempted to use several well-known methods of ML supervised and unsupervised machine learning algorithms, in order to examine the performance of each one accurately. Supervised learning algorithms made use of a decision tree (with the maximal depth of the decision tree being 10 and the minimal leaf size of the tree being 4. Also, the minimal size for split is 4), random forest, KNN, Gaussian Naïve Bayes, SVM, MLP (multilayer perceptron) and Linear Regression; whereas unsupervised learning algorithms applied K-means and DBSCAN algorithms. The test results for the supervised algorithms were as follows: decision tree with high F1-measure of more than 90% and with average precision, average recall and FP rate at respectively 90.92%, 90.5% and 0.062%, followed by the KNN algorithm with F1-measure of 89.04% and average precision, average recall and FP rate at respectively 90.09%, 89% and 0.0955%, had the best performance. The SVM algorithm with F1-measure of 74.5% and average precision, average recall and FP rate at respectively 76.92%, 70% and 0.117% had the worst performance. Also, the F1-measure of the results for the two unsupervised K-means and DBSCAN algorithms was obtained at around more than 50% (Figure 1). According to the results of this evaluation (Figure 2), the reduction in data scale led to a change in the behavior of the classifiers, so that the reduction of the property gradually increased the F1-measure of all classifiers. This upward trend in F1-measure continued up to 25 features, but after a certain value, the increasing trend stopped and the values of F1-measure decreased in some classes and remained fixed in some others. Two algorithms (KNN and decision tree) are among the most widely used models that will have different F1-measure values depending on the input data. In other words, the quality of data, scales and classes used in a dataset can affect the performance of each of the categories. In a dataset with a low number of inputs, KNN can increase the F1-measure of prediction due to using methods such as linear least squares approximation. This change in classifier behavior is clearly visible in the peaking phenomenon. According to this phenomenon, as the number of features increases from one point to the next, the classification error also increases. This phenomenon *per se* can indicate using proper value, for features (not too little and not too much). In this study, the appropriate number of features was 25, for which most methods yielded acceptable results.

The results of the study showed that unsupervised methods have poor performance for malware detection compared to supervised methods. This is not a good result to identify Android malware that is usually unknown. Although supervised learning is most often used, it requires that the outputs of the algorithm be already known and the data used to train the algorithm be already labeled with the correct answer. On the other hand, unsupervised machine learning is closer to what is called real artificial intelligence. Given the problem under investigation, which is the identification of malware on smartphones with increasing emergence of malicious samples, labeling all the data is not practically possible. Thus, it is better to use a method that can accurately identify unlabeled data, or, in other words, identify new ones. Hence, semi-supervised classification that has been on the path of development over the past few years is a good choice to identify unlabeled data, or in other words unknown malware, with high performance. In these methods, labeled data is trained by supervised methods and unlabeled data is grouped using a labeled dataset and labeled with the highest degree of assurance. Finally, all the labeled data is trained and evaluated using one of the classifiers.

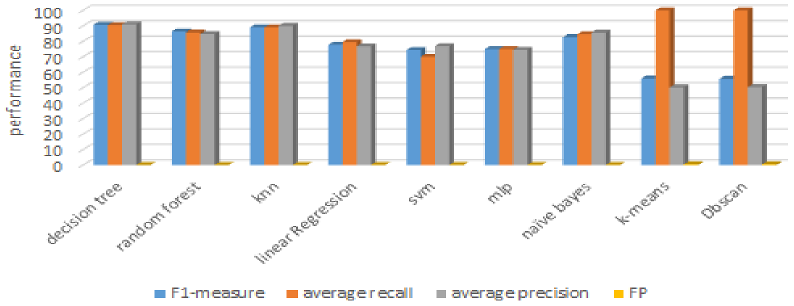


Figure 1. Results of the testing process with four criteria: F1-measure, average recall, average precision and FP.

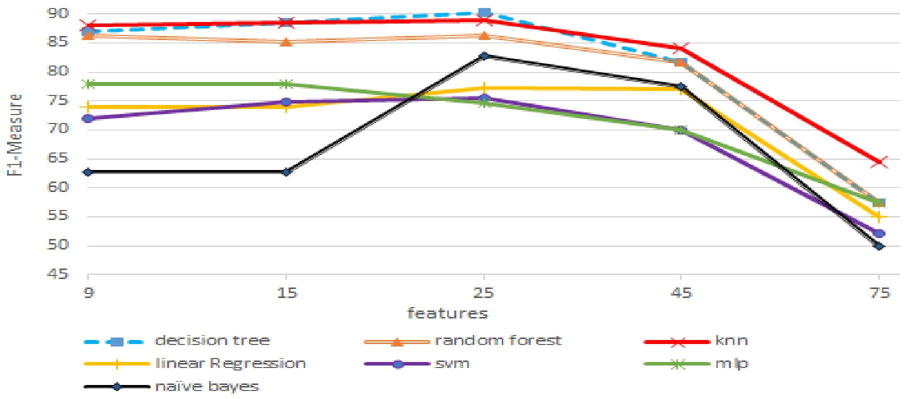


Figure 2. Benchmarking result in terms of the number of extracted features.

In the subsequent evaluations, we examined all experiments with 25 features as well as only the supervised algorithms.

Table 2. List of feature subsets.

Subsets	Features
9 features	F10, F11, F15, F16, F18, F42, F61, F69, F72
15 features	F10, F11, F13, F14, F15, F16, F18, F31, F42, F61, F69, F70, F71, F72, F74
25 features	F4, F5, F8, F9, F10, F11, F12, F13, F14, F15, F17, F18, F21, F25, F29, F30, F31, F42, F52, F61, F69, F70, F71, F72, F74
45 features	F1, F2, F3, F4, F5, F6, F8, F9, F10, F11, F12, F13, F14, F15, F16, F17, F18, F21, F25, F29, F30, F31, F32, F33, F34, F35, F36, F37, F38, F39, F42, F48, F49, F50, F51, F52, F53, F54, F55, F61, F69, F70, F71, F72, F74
75 features	F1,……, F75

4.2 Benchmarking in Terms of Time Duration of Capturing Traffic

The purpose of this evaluation is to assess the ability of various methods to detect malware in the shortest possible time. This issue is important, because in practice, if a system is supposed to detect malware, methods should be able to detect infections in a shorter time, as Internet access of phones to a connection point may not be possible for long. An applicable method should be able to detect malware as soon as possible by monitoring the traffic. In this benchmark, network traffic recording is marked every 15 minutes and classifiers are benchmarked using recorded traffic of 15, 30, 45 and 60 minutes, respectively. The results are shown in Figure 3. The F1-measure values of all methods increase with giving more traffic sampling time. The decision tree has a significant performance. Although KNN and decision tree are the best in 60 minutes of recorded traffic data, KNN cannot keep this performance for lesser time, while decision tree maintains this performance for 15 minutes of network traffic. The low level of F1-measure of other classifiers in the traffic volume of 15 minutes is due to the fact that many malwares have been widely generating traffic in more than 15 minutes of network presence. However, the decision tree has identified the presence of the first signals in the network.

This makes the decision tree the best choice in practice which can even detect a malware by recording smartphone traffic for a couple of minutes. The lowest F1-measure belongs to the Naïve Bayes method. This method had the least F1-measure in traffic for 15 minutes, while its F1-measure has improved significantly in traffic for 60 minutes compared to 15 minutes. Considering the need to rapid malware detection and the probability that a mobile phone would not be permanently connected to a long-term connectivity point, it seems that more research is needed to identify the appropriate features for high F1-measure detection in time durations as short as possible.

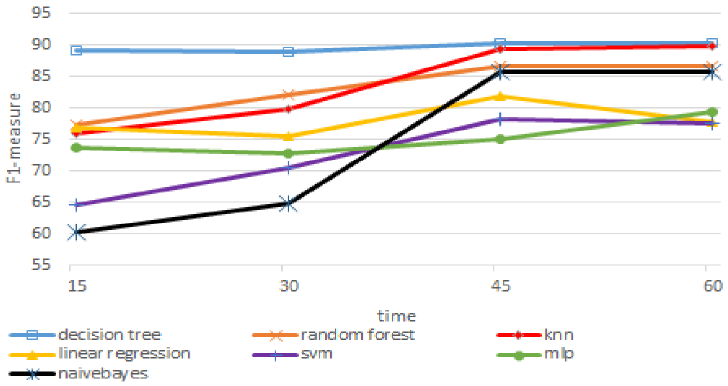


Figure 3. Benchmarking result in terms of the timing of program execution.

4.3 Benchmarking in Terms of the Ability to Detect Malware Families

Since malicious software can be categorized into the cybercrime group, it's important to identify types of malware families for proper response and defense after the attack. Identifying the behavioral patterns of traffic in each family can be very useful as a preventive measure of serious damage to Android devices. Therefore, in order to study the performance of machine learning classifications in identifying the type of malware family, two types of evaluation were performed. In the first one, the performance of each classifier was examined for identifying the type of malware family, while the second assessment evaluates the ability of each classifier to identify a new family of any malware that is separately evaluated in the next section.

In order to achieve better results, the experiment was performed with 25 attributes and a duration of 60 minutes. These were the best obtained adjustments in previous evaluations. In the first evaluation, four classes of malware families and 100 samples of them were tested. In this experiment, Naïve Bayes classification had a better performance with F1-measure of 74.83% compared to other classifiers. Decision tree classification had the lowest F1-measure (57.27%), (Figures 4-7).

Despite the best test conditions, the results were not satisfactory. The amount of data training is one of the important reasons for classification performance. In the case of the high number of classes, due to the lower number of data to be taught to each class, the classifier has less ability to distinguish between each class's patterns, which reduces the F1-measure of class identification. Other reasons include the type of input feature. Some of the features are noisy and they are not used in separating classes or some features cause incorrect recognition. The charts of F1-measure, average recall, average precision and FP values for each family are separately given in Figures 4-7. The high FP value indicates that the error rate of classification in identifying families from each other is significantly high. This can be confirmed for the F1-measure of this evaluation, based on the poor performance of classification in identifying families from one another. In this test, the features used to separate samples from each malware family have a lower ability of detaching malicious samples from benign samples.

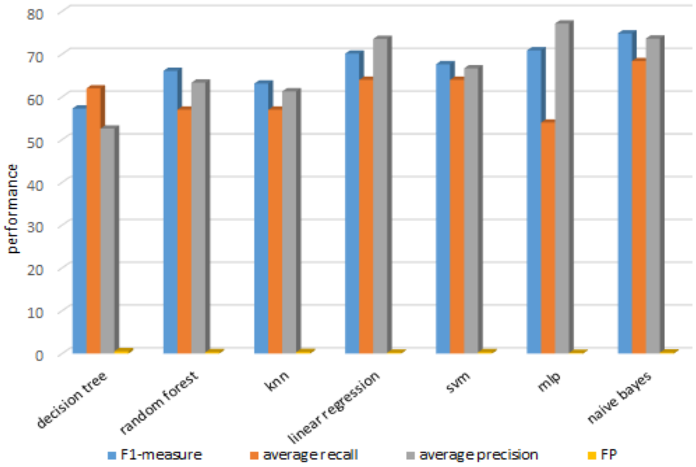


Figure 4. Benchmarking result in terms of the ability to identify Adware families.

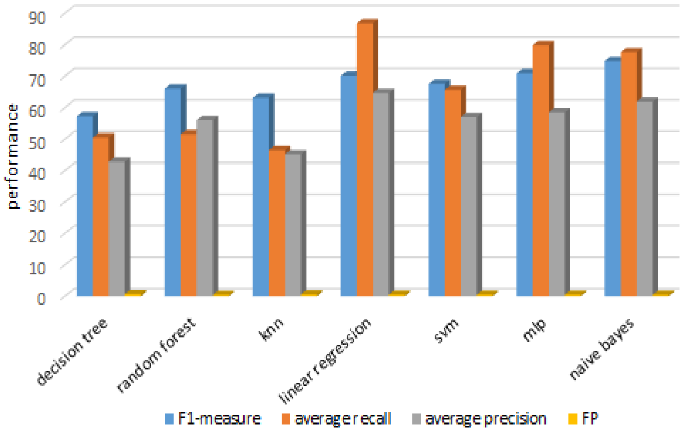


Figure 5. Benchmarking result in terms of the ability to identify Ransomware families.

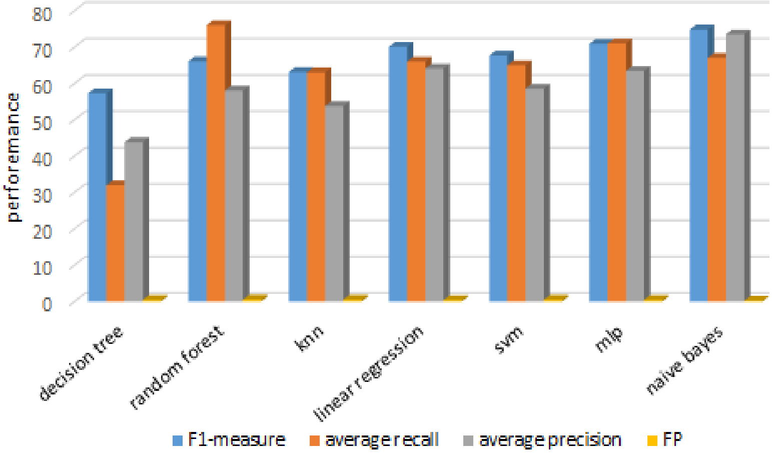


Figure 6. Benchmarking result in terms of the ability to identify Scareware families.

4.4 Benchmarking in Terms of the Ability to Identify Unknown Malware Types

In this evaluation, each new sample of families was separately examined. In this way, from each category of malware, a limited number of families were assigned to the test data and the rest of the families were used as the training data to the algorithm given. The purpose of this study is to measure the performance of machine learning classifications when new types of families are produced. These samples are initially unknown and each sample may have a different pattern. Therefore, the extraction of undiscovered patterns from unknowns and the association of these patterns with known samples can be easily accomplished by new methods such as machine learning. The results of the study showed that the F1-measure values obtained ranged from 21 to 80 percent, according to Figure 8. The SVM classifier achieved an F1-measure of over 80% and has been able to identify new samples of the Ransomware family. Also, the Naive Bayes classifier with the same F1-measure has been able to identify the family of SMS Malwares. However, the Random forest classifier with an average F1-measure of 66.49% had the best performance in identifying the four families and unexpectedly the KNN classifier with an average F1-measure of 54.92% in identifying the four families had the lowest level of detection F1-measure. Generally, classifier performance was better in identifying new samples in the Ransomware and SMS Malware families than in the families of Adware and Scareware. Since new families of

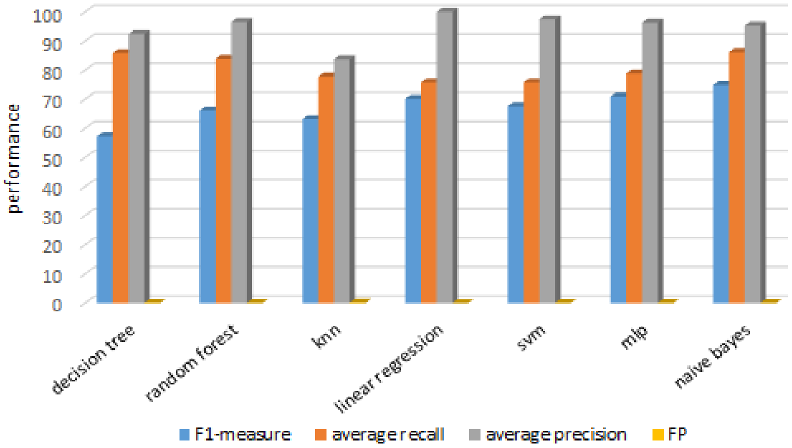


Figure 7. Benchmarking result in terms of the ability to identify SMS Malware families.

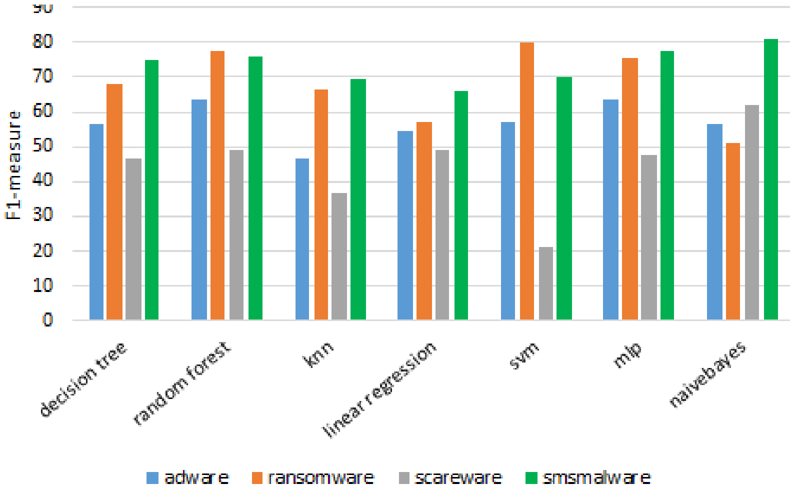


Figure 8. Benchmarking result in terms of the ability to identify new samples from each family.

Ransomware appear almost daily and traditional signature-based detection methods cannot be used to detect them, it is possible to combine deep system monitoring with machine learning, resulting in a system that can detect new families of Ransomware in real time by searching for certain behavioral patterns. Despite the fact that multiple families of Ransomware use different approaches to obfuscation, encryption and demand for ransom, the majority of them display similar behavioral patterns that can be detected. It can be argued that the traffic patterns of each family of malware are almost identical and that algorithms currently do not have the ability to identify each group of malware families. This low performance means the inability of the current algorithms to detect Android malware.

5. Limitations and Threats to Validity

The validity of benchmarking results may be threatened from several points which are mentioned and discussed below:

- Despite the title of the paper that is to benchmark Android malware detection machine learning algorithms, the studied methods are only able to detect those malwares which need to send packets over the network. These methods cannot be applicable if a malware does not send any packets on the network.
- The dataset studied in this paper has four categories of malware families; namely, Adware, Ransomware, Scareware and SMS Malware families. The authors do not claim that these results are valid for all malware families. Other case studies on malware detection may yield other results.
- The results represented in this paper are dedicated to CICAndMal2017 dataset which is a recently published one. The results for another dataset may be slightly different from those achieved in our study. For example, the current dataset includes the most popular applications. Obviously, any other sampling of applications, such as rarely used ones, requires a dedicated research.
- Although the dataset used in this paper is fresh, it is clear that changes in malware behavior in the future may lead to changes in the results.
- These results are in a situation where malwares are not aware of the existence of a malware detection system through network traffic.
- Malwares may bypass a detection mechanism by hiding or obfuscating their communications or by deceiving the machine learning algorithm through creating “adversarial examples” if they are aware of the existence of such a system.

6. Conclusions

One of the advantages of malware detection by analyzing network traffic is that through network traffic behavior, malicious samples can be identified before causing serious damage. In this paper, machine learning methods have been investigated in several aspects. These evaluations include the number of features needed to learn, the type of machine learning, the recorded traffic volume, the ability to identify the type of malware family and the ability to identify a new type of malware family. The investigated methods; namely, decision tree, Random forest, KNN, Linear Regression, SVM, MLP and Gaussian Naïve Bayes, were investigated in terms of sensitivity to the number of features examined. The results were as follows:

- The number of features should not be greater or less than a certain amount. Choosing the proper number of features and optimizing the size of the data lead to acceptable results.
- Almost all methods of increasing the time or increasing the volume of traffic recorded improved F1-measure. Only the decision tree algorithm was able to detect the presence of malware in a small volume of traffic with highly accurate prediction.
- In benchmarking the type of malware family, the performance of the algorithms was very low, considering the best conditions for the previous evaluations. Machine learning algorithms had a low ability to identify the behavior patterns of each family of malware.
- In terms of the last aspect benchmark to examine the performance of classification in identifying new or unknown samples. The results were also very low and not acceptable. It seems that further research is needed to identify the proper features for faster detection of malwares and their type at shorter time points.
- In all of the benchmarks, the type of data, the number and volume of data, as well as the number of noise features and the use of their proper number, have a great influence on the performance of all algorithms.

One of the main constraints of machine learning is the lack of prediction of points that are considered as noise. If a new data is received that belongs to the noise part, it can’t be classified using these methods.

Therefore, deep learning approaches at a deeper level than machine learning, inspired by the performance of the human brain and complex calculations on a large volume of data, solve issues thoroughly, having far better outcomes than machine learning approaches.

REFERENCES

[1] T. T. Mikko Hypponen, "F-Secure 2017 State of Cybersecurity Report," F-Secure, Tech. Rep., 2017.

[2] S. -H. Seo, A. Gupta, A. M. Sallam, E. Bertino and K. Yim, "Detecting Mobile Malware Threats to Homeland Security through Static Analysis," Journal of Network and Computer Applications, vol. 38, pp. 43-53, 2014.

[3] M. Finsterbusch, C. Richter, E. Rocha, J.-A. Muller and K. Hanssgen, "A Survey of Payload-based Traffic Classification Approaches," IEEE Communications Surveys & Tutorials, vol. 16, no. 2, pp. 1135-1156, 2013.

[4] H. Singh, "Performance Analysis of Unsupervised Machine Learning Techniques for Network Traffic Classification," Proc. of the 5th IEEE International Conference on Advanced Computing & Communication Technologies, pp. 401-404, , 2015.

[5] S. Zander, T. Nguyen and G. Armitage, "Automated Traffic Classification and Application Identification Using Machine Learning," Proc. of IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05), pp. 250-257, 2005.

[6] F. A. Narudin, A. Feizollah, N. B. Anuar and A. Gani, "Evaluation of Machine Learning Classifiers for Mobile Malware Detection," Soft Computing, vol. 20, no. 1, pp. 343-357, 2016.

[7] S. Garg, S. K. Peddoju and A. K. Sarje, "Network-based Detection of Android Malicious Apps,"International Journal of Information Security, vol. 16, no. 4, pp. 385-400, 2017.

[8] Y. Pang et al., "Finding Android Malware Trace from Highly Imbalanced Network Traffic," Proc. of IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), vol. 1, pp. 588-595, 2017.

[9] S. Pooryousef and K. Fouladi, "Proposing a New Feature for Structure-Aware Analysis of Android Malwares," Prco. of the 14th IEEE International ISC (Iranian Society of Cryptology) Conference on Information Security and Cryptology (ISCISC), pp. 93-98, 2017.

[10] A. Arora, S. Garg and S. K. Peddoju, "Malware Detection Using Network Traffic Analysis in Android Based Mobile Devices," Proc. of the 8th IEEE International Conference on Next Generation Mobile Apps, Services and Technologies, pp. 66-71, 2014.

[11] J. Erman, A. Mahanti and M. Arlitt, "Qrp05-4: Internet Traffic Identification Using Machine Learning," IEEE Globecom, pp. 1-6, 2006.

[12] J. Erman, M. Arlitt and A. Mahanti, "Traffic Classification Using Clustering Algorithms," Proceedings of the 2006 SIGCOMM Workshop on Mining Network Data, ACM, pp. 281-286, 2006.

[13] A. Arora and S. K. Peddoju, "Minimizing Network Traffic Features for Android Mobile Malware Detection," Proceedings of the 18th International Conference on Distributed Computing and Networking, ACM, p. 32, 2017.

[14] T. Bujlow, T. Riaz and J. M. Pedersen, "Classification of HTTP Traffic Based on C5. 0 Machine Learning Algorithm," Proc. of IEEE Symposium on Computers and Communications (ISCC), pp. 000882-000887, 2012.

[15] O. M. Alhawi, J. Baldwin and A. Dehghantanha, "Leveraging Machine Learning Techniques for Windows Ransomware Network Traffic Detection," Cyber Threat Intelligence, pp. 93-106, 2018.

[16] F. Pendlebury, F. Pierazzi, R. Jordaney, J. Kinder and L. Cavallaro, "TESSERACT: Eliminating Experimental Bias in Malware Classification across Space and Time," arXiv preprint arXiv:1807.07838, 2018.

[17] D. Nancy and D. Sharma, "Android Malware Detection Using Decision Trees and Network Traffic," International Journal of Computer Science and Information Technologies, vol. 7, no. 4, pp. 1970-1974, 2016.

[18] Z. Chen et al., "Machine Learning Based Mobile Malware Detection Using Highly Imbalanced Network Traffic," Information Sciences, vol. 433, pp. 346-364, 2018.

[19] M. Zaman, T. Siddiqui, M. R. Amin and M. S. Hossain, "Malware Detection in Android by Network Traffic Analysis," Proc. of IEEE International Conference on Networking Systems and Security (NSysS), pp. 1-5, 2015.

[20] K. Allix, T. F. D. A. Bissyande, Q. Jerome, J. Klein and Y. Le Traon, "Empirical Assessment of Machine Learning-based Malware Detectors for Android: Measuring the Gap Between in-the-lab and in-the-wild Validation Scenarios," Empirical Software Engineering, pp. 1-29, 2014.

[21] L. Chen, M. Zhang, C.-Y. Yang and R. Sahita, "Semi-supervised Classification for Dynamic Android Malware Detection," arXiv preprint arXiv:1704.05948, 2017.

[22] H. Debar, M. Dacier and A. Wespi, "A Revised Taxonomy for Intrusion-detection Systems," Annales des Télécommunications, Springer, vol. 55, no. 7-8, pp. 361-378, 2000.

[23] I. Homoliak, Intrusion Detection in Network Traffic, Dissertation, Faculty of Information Technology, University of Technology, 2016.

[24] A. H. Lashkari, A. F. A. Kadir, H. Gonzalez, K. F. Mbah and A. A. Ghorbani, "Towards a Network-based Framework for Android Malware Detection and Characterization," Proc. of the 15th IEEE Annual Conference on Privacy, Security and Trust (PST), pp. 233-239, 2017.

[25] A. H. Lashkari, A. F. A. Kadir, L. Taheri and A. A. Ghorbani, "Toward Developing a Systematic Approach to Generate Benchmark Android Malware Datasets and Classification," Proc. of IEEE International Carnahan Conference on Security Technology (ICCST), pp. 1-7, 2018.

[26] A. Jović, K. Brkić and N. Bogunović, "A Review of Feature Selection Methods with Applications," Proc. of the 38th IEEE International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1200-1205, 2015.

[27] S. Maldonado, R. Weber and J. Basak, "Simultaneous Feature Selection and Classification Using Kernel-penalized Support Vector Machines," Information Sciences, vol. 181, no. 1, pp. 115-128, 2011.

[28] M. Q. Nguyen and J. P. Allebach, "Feature Ranking and Selection Used in a Machine Learning Framework for Predicting Uniformity of Printed Pages," Electronic Imaging, vol. 2017, no. 12, pp. 166-173, 2017.

[29] H. Alaidaros, M. Mahmuddin and A. Al Mazari, "An Overview of Flow-based and Packet-based Intrusion Detection Performance in High Speed Networks," Proceedings of the International Arab Conference on Information Technology, pp. 1-9, 2011.

ملخص البحث:

أصبحت الهواتف الذكية تحتل جزءاً أساسياً من حياة الناس هذه الأيام، وقد قاد ذلك الى انخراط أعداد متزايدة من المستخدمين في استخدام هذه التقنية. ولقد شجع ازدياد أعداد المستخدمين القراصنة الإلكترونيين على إنشاء تطبيقات هدفها إلحاق الأذى بمستخدمي الهواتف الذكية. وما من شك في أن تحديد الاختراقات التي يتعرض لها المستخدمون والكشف عنها أمر حاسم من أجل الحفاظ على أمن المستخدمين وخصوصيتهم. وتبين الاتجاهات الحديثة للأمن السيبراني أنه يمكن تحديد تلك التهديدات بصورة فعالة باستخدام تقنيات كشف ترتكز على الشبكة وعدد من طرق تعلم الآلة.

وفي هذه الورقة، تم استقصاء عدد من طرق تعلم الآلة المعروفة جيداًمن أجل الكشف عن الاختراقات التي تتعرض لها الهواتف الذكية باستخدام حركة المرور على الشبكة. واستخدمت عدة عائلات اختراق في الاستقصاء (أدوير "Adware"؛ رنسم وير "Ransomware"؛ سكيروير "Scareware"؛ إس إم إس مالوير "SMS Malware"). من جهة أخرى، جرى استخدام طرق تعلم الآلة الأوسع استخداماً والأكثر شهرة؛ المراقبة وغير المراقبة. وقارنت الدراسة بين هذه الطرق من عدة جوانب، مثل عدد السِّمات المطلوبة، وحركة المرور المسجلة، والقدرة على كشف عائلة الاختراق، والقدرة على كشف عائلة اختراق جديدة.

وبينت النتائج أن استخدام هذه الطرق بالسمات المناسبة وحركة المرور ذات الحجم المناسب من شأنه أن يحقق نسبة كشف اختراقات تقرب من 90%. غير أن هذه الطرق لم تحقق نتائج مقبولة من حيث كشف عائلات الاختراق أو عائلات الاختراق الجديدة غير المعروفة. ومن جهة اخرى، شرحت الدراسة بعض التحديات ووجهت الى مسائل محتملة يمكن للباحثين أن يتناولوها مستقبلاً بناءً على نتائج الدراسة الحالية.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 05, No. 03, December 2019.

3-D Polarized Channel Modeling for Multipolarized UCA-Massive MIMO Systems in Uplink Transmission

Abdelhamid Riadi[1], Mohamed Boulouird[1, 2] and Moha M'Rabet Hassani[1]

(Received: 20-Jul.-2019, Revised: 23-Sep.-2019, Accepted: 8-Oct.-2019)

Abstract

In this paper, a novel design of a Uniform Circular Array Massive-Multiple Input Multiple Output (UCA-mMIMO) system based on Spherical Wave (SW) is proposed in Uplink (UL) transmission. A three-dimensional (3-D) channel pattern is established and estimated, where channel orthogonality of multipolarized/unipolarized UCA-mMIMO systems is analyzed. Multipolarized and unipolarized systems are evaluated to decrease channel orthogonality. The Azimuth Angle of Arrival (AAoA) and Elevation Angle of Arrival (EAoA), as well as antenna spacing and cross-polarization discrimination, are taken into consideration. Using Monte Carlo simulation method, the results show that the multipolarized UCA-mMIMO system provides a better performance compared to the unipolarized UCA-mMIMO system in different situations. The proposed design is homely to be realized in real environment in conformance to the parameters analyzed; in order to confirm that it will be a very good choice.

Keywords

Massive MIMO system, Channel orthogonality, Channel estimation, Multipolarized, Unipolarized, OSIC detector.

1. Introduction

Massive MIMO system is one of the successful technologies for the new-generation 5G. High-quality communication represented in features such as voice, audiovisual communication …etc., is promoted by using mMIMO systems, in addition to that the growing number of terminals requires a high throughput [1]. Furthermore, several publications have appeared in recent years focusing on spectral efficiency enhancement in wireless communication [2]-[4]; collecting mMIMO with Orthogonal Frequency Division Multiplexing (OFDM) can support high spectrum efficiency [5]. In the same way, the compromise between energy efficiency and spectral efficiency; measured in terms of (bits/j) and (bits/channel use/terminal), respectively, is derived using the convex optimization theory [4], where this trade-off is quantified in the case of a channel model that contains small-scale fading. Otherwise, the classical MIMO (i.e., 4G) system has a tendency to use four or eight antennas, while in mMIMO system, especially in a single cell, Base Station (BS) antennas are larger than several terminals [1]-[2], [6]. Additionally, the channel between the transceiver is an important element in mMIMO system. Due to various phenomena, such as diffraction, interference, reflection, …etc., the system performance is degraded. In the past decade, a lot of research has attracted attention to investigate perfect/imperfect Channel State Information (CSI) [7]. In the same way, when the CSI is unfinished at the BS antennas, the data detection contains erroneous bits and the system performance is deteriorated. Accordingly, various works focused on investigating the channel estimation phenomenon, in which Least-Square Channel Estimation (LSCE) is used for its simplicity and low complexity [1], [8], on the one hand. On the other hand, linear detectors, such as Zero Forcing (ZF) and Minimum Mean-Square Error (MMSE), are widely used to detect the stream data [1], [8]-[11]. Moreover, Ordered Successive Interference Cancellation (OSIC) is generally better than simple linear detectors (i.e., ZF, MMSE) [1]. Therefore, OSIC is evaluated under various criteria, such as declining Signal-to-Noise-Ratio (SNR) criterion [12] and greatest SNR criterion [13]. In addition to that, favorable propagation (i.e., channel orthogonality) is one of leading properties in mMIMO system [1]. Moreover, low channel orthogonality for different 3-D channel models has been investigated with Uniform Linear Array

1. A. Riadi, M. Hassani and M. Boulouird are with Instrumentation, Signals and Physical Systems (I2SP), Faculty of Sciences Semlalia Cadi Ayyad University, Marrakech, Morocco, E-mail: abdelhamid.riadi@edu.uca.ac.ma

2. M. Boulouird is with National School of Applied Sciences of Marrakech (ENSA-M), Cadi Ayyad University, Marrakech, Morocco, E-mail: m.boulouird@uca.ac.ma

(ULA), Uniform Rectangular Array (URA) and UCA-mMIMO using Plane Wave (PW) in many situations [14]. Similarly, PW and SW are discussed for 3-D ULA m-MIMO [15]. Furthermore, in this work, our contributions are summarized as follows:

- A new geometrical conception is realized for multipolarized UCA-mMIMO system using SW;
- A 3-D channel pattern with various parameters is modeled and estimated.

The remainder of the paper is organized as follows. In Section 2, the mMIMO model is presented in Uplink (UL) transmission, where single cell is considered. Section 3 describes the outlines of LSCE. Channel modeling for UCA-mMIMO using SW is evaluated for both multipolarized/unipolarized antennas in Section 4. In Section 5, the OSIC detector is discussed based on ZF and MMSE detectors. Section 6 presents the simulation and analysis results. Section 7 summarizes the results of this paper and draws conclusions.

2. Massive MIMO Model

In this section, a Massive-MIMO-OFDM system is considered in Uplink transmission from N_t terminals with single antennas to a single BS with N_r antennas. The studied system is gived in Figure 1. The length of sub-carriers and the cyclic prefix (CP) are defined by K and ν, respectively. The CP is inserted on each transmitting antenna to achieve a full OFDM symbol. In this paper, the CP is superior to the utmost multi-path delay [1], [8] and [16]-[17].

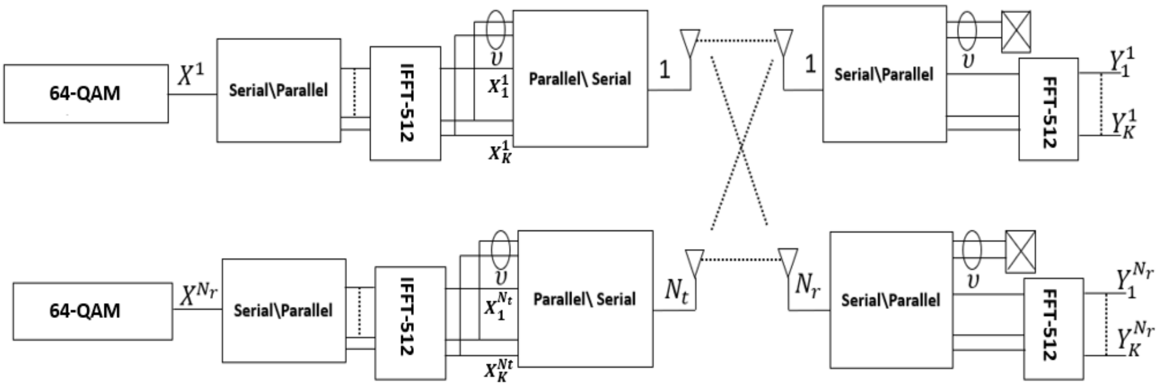


Figure 1. System model.

In the same way, at the reception side, the CP is removed on each receiving antenna. Taking-for example-the q^{th} receiving antenna, the received signal vector $y^q(n)$ is K×1 and can be expressed as follows:

$$y^q(n) = \sum_{r=1}^{N_t} H_{cir}^{q,r} F^H X^r(n) + z^q(n) \tag{1}$$

From Equation 1, the circulant matrix $H_{cir}^{q,r}$ has a first column defined by $[\mathfrak{h}^{q,r^T}, 0_{1\times(K-L)}]^T$, in addition to that L is the length of the channel impulse response and $\mathfrak{h}^{q,r}$ represents the L×1 vector. The OFDM vector that is transmitted on each transmitting antenna is defined by $X^r(n)$ with K×1 dimensions, r and n are indices of number of transmitting antenna and time, respectively. $z^q(n)$ is the additive Gaussian noise at Time Index (TI) n with zero mean and a variance of σ_n^2. Moreover, the unitary DFT matrix with dimensions K×K is presented by F. From the eigenvalue decomposition, the circulant matrix becomes $H_{cir}^{q,r} = F^H diag\{\sqrt{K} F [\mathfrak{h}^{q,r^T}, 0_{1\times(K-L)}]^T\} F$ [1], [8] and [17]. Finally, the FFT of the received signal $y^q(n)$ is given as follows:

$$Y^q(n) = \sum_{r=1}^{N_t} diag\{\sqrt{K} F [\mathfrak{h}^{q,r^T}, 0_{1\times(K-L)}]^T\} \times X^r(n) + \Xi^q(n) \tag{2}$$

where, $\Xi^q(n) = F z^q(n)$.

3. Least Square Massive MIMO

Based on the same system presented in Figure 1, the LSCE scheme is presented. Then, Equation 2 can be written as:

$$\begin{aligned} Y^q(n) &= \textstyle\sum_{r=1}^{N_t} diag\{X^r(n)\}F\mathfrak{h}^{q,r} + \Xi^q(n) \\ &= \textstyle\sum_{r=1}^{N_t} (diag\{D^r(n)\} + diag\{B^r(n)\})F\mathfrak{h}^{q,r} + \Xi^q(n) \end{aligned} \tag{3}$$

From Equation 3, F is $\sqrt{K} \times l$ of F, where l is the 1st column of F, noting that $D_{diag}^r(n) = diag\{D^r(n)\}$ and $B_{diag}^r(n) = diag\{B^r(n)\}$, where $D_{diag}^r(n)$ and $B_{diag}^r(n)$ are $K \times 1$ data vector and $K \times 1$ pilot sequence vector, respectively. Hence, Equation 3 becomes:

$$Y^q(n) = \textstyle\sum_{r=1}^{Nt} D_{diag}^r(n)F\mathfrak{h}^{q,r} + \sum_{r=1}^{Nt} B_{diag}^r(n)F\mathfrak{h}^{q,r} + \Xi^q(n) \tag{4}$$

Furthermore, in this work, the training of all OFDM symbols is done at maximum value of g and TI is $n \in \{0, \cdots, g-1\}$ [1], [8]. We consider the data model:

$$Y^q = T\mathfrak{h}^q + A\mathfrak{h}^q + \Xi^q \tag{5}$$

where, $Y^q = [Y^{q^T}(0), \cdots, Y^{q^T}(g-1)]^T$, $\Xi^q = [\Xi^{q^T}(0), \cdots, \Xi^{q^T}(g-1)]^T$,

$$A = \begin{bmatrix} B_{diag}^1(0)F & \cdots & B_{diag}^{Nt}(0)F \\ \vdots & & \vdots \\ B_{diag}^1(g-1)F & \cdots & B_{diag}^{Nt}(g-1)F \end{bmatrix} T = \begin{bmatrix} D_{diag}^1(0)F & \cdots & D_{diag}^{Nt}(0)F \\ \vdots & & \vdots \\ D_{diag}^1(g-1)F & \cdots & D_{diag}^{Nt}(g-1)F \end{bmatrix} \tag{6}$$

and $\mathfrak{h}^q = [\mathfrak{h}^{q,1^T}, \cdots, \mathfrak{h}^{q,Nt^T}]^T$.

The LSCE technique minimizes the noise defined in Equation 5 [1], [8], based on the cost function (Equation 7), to obtain the estimated channel noted by $\hat{\mathfrak{h}}^q$

$$\begin{aligned} J(\hat{\mathfrak{h}}^q) &= ||Y^q - A\hat{\mathfrak{h}}^q||^2 \\ &= (Y^q - A\hat{\mathfrak{h}}^q)^H(Y^q - A\hat{\mathfrak{h}}^q) \\ &= Y^{q^H}Y^q - Y^{q^H}A\hat{\mathfrak{h}}^q - \hat{\mathfrak{h}}^{q^H}A^HY^q + \hat{\mathfrak{h}}^{q^H}A^HA\hat{\mathfrak{h}}^q \end{aligned} \tag{7}$$

Next, we take the derivation of Equation 7 relative to $\hat{\mathfrak{h}}^q$ variable,

$$\frac{\partial J(\hat{\mathfrak{h}}^q)}{\partial \hat{\mathfrak{h}}^q} = 2 * (-(A^HY^q)^* + (A^HA\hat{\mathfrak{h}}^q)^*) = 0 \tag{8}$$

Finally, we have $A^HA\hat{\mathfrak{h}}^q = A^HY^q$ and the solution of the LSCE is given by the following expression:

$$\hat{\mathfrak{h}}^q = A^+Y^q \tag{9}$$

where, A^+ is the pseudo-inverse that is equal to $(A^HA)^{-1}A^H$ if $gK \geqslant LN_t$. Because $rank(A) = min(gK, LN_t)$, the necessary and sufficient condition to have unique LSCE is $gK \geqslant LN_t$. This LS method presents low complexity and high simplicity. In addition, taking the information about the channel and the noise is not necessary [1], [8] and [17]-[19]. From Equation 5, we then find that:

$$\hat{\mathfrak{h}}^q = \mathfrak{h}^q + A^+T\mathfrak{h}^q + A^+\Xi^q \tag{10}$$

Further, to suppress the interference due to the data, the following condition is imposed:

$$A^+T = 0_{LN_t \times LN_t} \tag{11}$$

Furthermore, satisfying this condition requires choosing disjoint sets of pilot tones for training and data in each OFDM symbol (i.e., zeros in $B^r(n)$, where $D^r(n)$ contains non-zeros and inversely). Equation 10 then becomes:

$$\hat{\mathfrak{h}}^q = \mathfrak{h}^q + A^+ \Xi^q \tag{12}$$

Equation 12 is an association of two parts; the first is the true channel $\mathfrak{h}^q$ and the second is the noise in the system. Thus, for zero-mean noise, $\varepsilon\{\hat{\mathfrak{h}}^q\} = \mathfrak{h}^q + A^+ \varepsilon\{\Xi^q\} = \mathfrak{h}^q$, (i.e., $\hat{\mathfrak{h}}^q$ forms an unbiased estimate of $\mathfrak{h}^q$). Furthermore, the estimated channel matrix $\hat{\mathbb{H}} \in \mathbb{C}^{N_r \times N_t}$ which includes all terminal antennas N_t and all BS antennas N_r is given by:

$$\hat{\mathbb{H}} = \begin{bmatrix} \hat{\mathfrak{h}}^{1,1} & \cdots & \hat{\mathfrak{h}}^{1,N_t} \\ \vdots & & \vdots \\ \hat{\mathfrak{h}}^{q,1} & \cdots & \hat{\mathfrak{h}}^{q,N_t} \\ \vdots & & \vdots \\ \hat{\mathfrak{h}}^{N_r,1} & \cdots & \hat{\mathfrak{h}}^{N_r,N_t} \end{bmatrix} \tag{13}$$

where the estimated channel vector at terminal position i is given by $\hat{\mathcal{H}}_i = [\hat{\mathfrak{h}}^{1,i^T}, \cdots, \hat{\mathfrak{h}}^{N_r,i^T}]^T$.

4. Channel Modeling

In this section, the UCA-mMIMO system based on SW is investigated as shown in Figure 2. Form this configuration, the horizontally polarized antenna A_1 is considered as a reference. In addition to that, all the odd ciphered antennas are horizontally polarized antennas and all the even ciphered antennas are vertically polarized antennas. Note that d is the distance between two adjacent antennas. From Figure 2, $\theta = \frac{2\pi}{N_r}$, $\theta_2 = \theta$, $\theta_3 = 2\theta$, $\cdots$, $\theta_m = (m-1)\theta$ and the radius r is equal to $\frac{d}{2sin(\frac{\theta}{2})}$. Similarly, the signals arrive from the S location; also, $S_1, S_2, \cdots, S_m$ are successive projections of S on the horizontal planes $x - A_1 - y, x - A_2 - y, \cdots, x - A_m - y$, respectively. Each projection has a distance from the source S noted by $h = h_1, h_2, \cdots, h_m$, respectively. $d_{x_1}, d_{x_2}, \cdots, d_{x_m}$ are the projections on the x axis of $S_1, S_2, \cdots, S_m$, respectively, while d_y denotes their projections on the y axis.

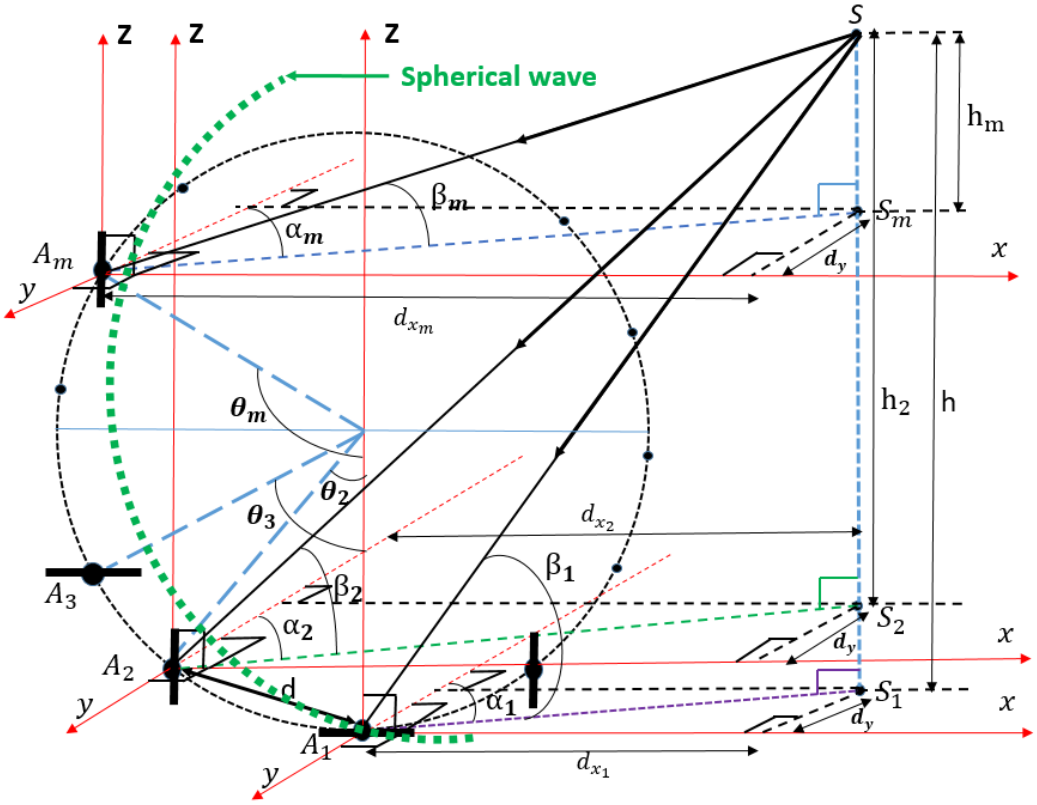


Figure 2. Multipolarized uniform circular array Massive-MIMO configuration.

Otherwise, the distance between each antenna and the source S can be defined by:

$$d_{SA_1} = \sqrt{d_{x_1}^2 + d_y^2 + h_1^2} \tag{14}$$

$$d_{SA_2} = \sqrt{(d_{x_1} + r * sin(\theta))^2 + d_y^2 + h_2^2} \tag{15}$$

$$d_{SA_m} = \sqrt{(d_{x_1} + r * sin((m-1)\theta))^2 + d_y^2 + h_m^2} \tag{16}$$

$$d_{SA_{N_r}} = \sqrt{(d_{x_1} + r * sin((N_r - 1)\theta))^2 + d_y^2 + h_{N_r}^2} \tag{17}$$

Depending on the geometrical relationship presented in Figure (2), for an arbitrary antenna, we have:

$$d_{x_1} = d_y * tan(\alpha_1) \tag{18}$$

$$h_m = tan(\beta_m) * \sqrt{d_y^2 + d_y^2 * tan(\alpha_m)^2} \tag{19}$$

In addition to that, the antenna A_1 is considered as a reference antenna. The estimated path of multipath channels at terminal position i and at BS antennas A_1 and A_m using SW can be expressed by:

$$\hat{\mathfrak{h}}_{UCA}^{1,i} = \sqrt{P_H} e^{j(\phi_i + 2\pi \sqrt{d_{x_{1,i}}^2 + d_{y,i}^2 + h_{1,i}^2 / \lambda)}} \tag{20}$$

$$\hat{\mathfrak{h}}_{UCA}^{m,i} = \sqrt{P_V} e^{j(\phi_i + 2\pi \sqrt{(d_{x_{1,i}} + r * sin((m-1)\theta))^2 + d_{y,i}^2 + h_{m,i}^2 / \lambda)}} \tag{21}$$

where, P_H and P_V are the horizontally polarized power and vertically polarized power in this path, respectively. λ is the wavelength and ϕ is the uniform random phase as assumed by IID on $[-\pi, \pi)$ [14].

Otherwise, the uniform distribution of AAS and EAS is used to describe the AAoA and EAoA distribution. In this paper, α and β represent AAS and EAS, respectively; several authors [14], [20] have established the Power Azimuth Spread (PAS) to explain the AAoA/EAoA as follows:

$$p(\Upsilon) = \frac{1}{2\Delta\Upsilon}, \quad -\Delta\Upsilon + \Upsilon_0 \leqslant \Upsilon \leqslant \Delta\Upsilon + \Upsilon_0, \tag{22}$$

where, Υ_0 and $\Delta\Upsilon$ are the mean of AAoA/EAoA and AAS/EAS, respectively.

Moreover, in real environment and thanks to the multipath phenomenon, polarization can be changed between the terminal and the BS antennas. Thereby, to characterize this phenomenon, cross-polarization discrimination (XPD) is expressed as [14], [25]:

$$XPD = \frac{E\{|\mathfrak{h}_{VV}|^2\}}{E\{|\mathfrak{h}_{VH}|^2\}} = \frac{E\{|\mathfrak{h}_{HH}|^2\}}{E\{|\mathfrak{h}_{HV}|^2\}} = \frac{1-a}{a}, \tag{23}$$

where, $E\{\}$ describes the expectation operator, $\mathfrak{h}_{VV}/\mathfrak{h}_{HH}$ and $\mathfrak{h}_{VH}/\mathfrak{h}_{HV}$ are the channels between transceivers with the same polarization and with different polarizations, respectively. Then, Equations 20 and 21 can be written as:

$$\hat{\mathfrak{h}}_{UCA,XPD}^{1,i} = \sqrt{P_V a + P_H(1-a)} e^{j(\phi_i + 2\pi \sqrt{d_{x_{1,i}}^2 + d_{y,i}^2 + h_{1,i}^2 / \lambda)}} \tag{24}$$

$$\hat{\mathfrak{h}}_{UCA,XPD}^{m,i} = \sqrt{P_H a + P_V(1-a)} e^{j(\phi_i + 2\pi \sqrt{(d_{x_{1,i}} + r * sin((m-1)\theta))^2 + d_{y,i}^2 + h_{m,i}^2 / \lambda)}} \tag{25}$$

Furthermore, in this paper and in related references [14], [21]-[24], $a (0 < a \leq 1)$ represents the power that is seeped in the UL transmission from the vertically/horizontally polarized terminal to the horizontally/vertically polarized BS antenna; in the case where there is no seepage, a is equal to 0. In addition to that, $P_X(1-a)$ is the power kept in similar transmission and reception polarization states, where $P_X a$ is the power seepage to transmission in one polarization state and to reception in the orthogonal polarization state ($X \in V, H$).

From our proposed configuration, Equations 24 and 25 can be generalized for all multipaths at terminal position e with single horizontally polarized antenna side and f with single vertically polarized antenna side, over all BS antennas (i.e., $\{1, \cdots, N_r\}$) basing on SW by:

$$\hat{\mathcal{H}}_e^{UCA,XPD} = \begin{bmatrix} \hat{\mathfrak{h}}_{UCA,XPD}^{1,e} \\ \hat{\mathfrak{h}}_{UCA,XPD}^{2,e} \\ \vdots \\ \hat{\mathfrak{h}}_{UCA,XPD}^{m,e} \\ \hat{\mathfrak{h}}_{UCA,XPD}^{m+1,e} \\ \vdots \\ \hat{\mathfrak{h}}_{UCA,XPD}^{N_r,e} \end{bmatrix} = \begin{bmatrix} \sqrt{P_H(1-a)}e^{j(\phi_e+2\pi\sqrt{d_{x_1,e}^2+d_{y,e}^2+h_{1,e}^2/\lambda)}} \\ \sqrt{P_H(a)}e^{j(\phi_e+2\pi\sqrt{(d_{x_1,e}+r*sin(\theta))^2+d_{y,e}^2+h_{2,e}^2/\lambda)}} \\ \vdots \\ \sqrt{P_H(a)}e^{j(\phi_e+2\pi\sqrt{(d_{x_1,e}+r*sin((m-1)\theta))^2+d_{y,e}^2+h_{m,e}^2/\lambda)}} \\ \sqrt{P_H(1-a)}e^{j(\phi_e+2\pi\sqrt{(d_{x_1,e}+r*sin((m)\theta))^2+d_{y,e}^2+h_{m+1,e}^2/\lambda)}} \\ \vdots \\ \sqrt{P_H(a)}e^{j(\phi_e+2\pi\sqrt{(d_{x_1,e}+r*sin((N_r-1)\theta))^2+d_{y,e}^2+h_{N_r,e}^2/\lambda)}} \end{bmatrix}. \quad (26)$$

$$\hat{\mathcal{H}}_f^{UCA,XPD} = \begin{bmatrix} \hat{\mathfrak{h}}_{UCA,XPD}^{1,f} \\ \hat{\mathfrak{h}}_{UCA,XPD}^{2,f} \\ \vdots \\ \hat{\mathfrak{h}}_{UCA,XPD}^{m,f} \\ \hat{\mathfrak{h}}_{UCA,XPD}^{m+1,f} \\ \vdots \\ \hat{\mathfrak{h}}_{UCA,XPD}^{N_r,f} \end{bmatrix} = \begin{bmatrix} \sqrt{P_V(a)}e^{j(\phi_f+2\pi\sqrt{d_{x_1,f}^2+d_{y,f}^2+h_{1,f}^2/\lambda)}} \\ \sqrt{P_V(1-a)}e^{j(\phi_f+2\pi\sqrt{(d_{x_1,f}+r*sin(\theta))^2+d_{y,f}^2+h_{2,f}^2/\lambda)}} \\ \vdots \\ \sqrt{P_V(1-a)}e^{j(\phi_f+2\pi\sqrt{(d_{x_1,f}+r*sin((m-1)\theta))^2+d_{y,f}^2+h_{m,f}^2/\lambda)}} \\ \sqrt{P_V(a)}e^{j(\phi_f+2\pi\sqrt{(d_{x_1,f}+r*sin((m)\theta))^2+d_{y,f}^2+h_{m+1,f}^2/\lambda)}} \\ \vdots \\ \sqrt{P_V(1-a)}e^{j(\phi_f+2\pi\sqrt{(d_{x_1,f}+r*sin((N_r-1)\theta))^2+d_{y,f}^2+h_{N_r,f}^2/\lambda)}} \end{bmatrix}. \quad (27)$$

From Equations 26 and 27, P_H and P_V are the horizontally and vertically polarized power of terminals e and f, respectively. In this paper, after getting the estimated channel vectors of terminal e with single horizontally polarized antenna and f with vertically polarized antenna for UCA-mMIMO systems in Equations 26 and 27, the channel orthogonality between $\hat{\mathcal{H}}_e^{UCA,XPD}$ and $\hat{\mathcal{H}}_f^{UCA,XPD}$ is defined by [14], [26]-[27]:

$$\delta_{e,f} = \frac{|(\hat{\mathcal{H}}_e^{UCA,XPD})^H \hat{\mathcal{H}}_f^{UCA,XPD}|}{||\hat{\mathcal{H}}_e^{UCA,XPD}||.||\hat{\mathcal{H}}_f^{UCA,XPD}||} \quad (28)$$

where, $||.||$ represents the Euclidean norm. Similarly, the estimated channel matrix (13) can be rewritten as $\hat{\mathbb{H}} = [\hat{\mathcal{H}}_1^{UCA,XPD}, \cdots, \hat{\mathcal{H}}_e^{UCA,XPD}, \cdots, \hat{\mathcal{H}}_f^{UCA,XPD}, \cdots, \hat{\mathcal{H}}_{N_t}^{UCA,XPD}]$; where, the estimated vector at terminal position i is noted by $\hat{\mathcal{H}}_i^{UCA,XPD} = [\hat{\mathfrak{h}}_{UCA,XPD}^{1,i^T}, \cdots, \hat{\mathfrak{h}}_{UCA,XPD}^{N_r,i^T}]^T$. In the same way, according to literature, the authors of [14] and [28] supposed that a half of the terminals are horizontally polarized antennas and the other half of the terminals are vertically polarized antennas at a short time due to arbitrary location.

5. OSIC Signal Detection

In this part of the paper, we present an important class of nonlinear signal detection, specifically the OSIC [1], [8]-[9]. Hence, according to Figure 3, we illustrate the OSIC signal detection for an example of three spatial streams. Furthermore, for symbol detection, the Linear Transformation Matrix (LTM), defined as $T_{ZF} = (\widehat{\mathbb{H}}^H \widehat{\mathbb{H}})^{-1} \widehat{\mathbb{H}}^H$ [9]-[11], [29], is taken into account. Moreover, the first data is detected with the first row vector of LTM (i.e., T_{ZF}); after the slicing process is carried out, x_1 is created. The interference due to the detected stream in the first stage is subtracted from the received signal; that is $y_1 = y - \widehat{\mathcal{H}}_1^{UCA,XPD} x_1$, where $y = [Y^1, \cdots, Y^q, \cdots, Y^{N_r}]^T$. Hence, the interference from the first stage is canceled; in addition to that, another stream is detected and sliced in the second stage x_2 and the interference is canceled by $y_2 = y_1 - \widehat{\mathcal{H}}_2^{UCA,XPD} x_2$. In the same way, detection and slicing of the stream as well as the interference cancellation steps are carried out in each stage [1], [8] and [29]-[32].

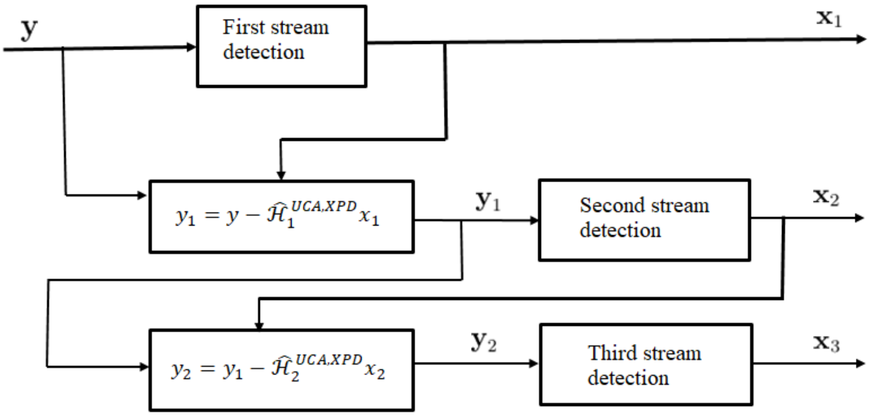

y
First stream detection
x1
$y_1 = y - \widehat{\mathcal{H}}_1^{UCA,XPD} x_1$
y1
Second stream detection
x2
$y_2 = y_1 - \widehat{\mathcal{H}}_2^{UCA,XPD} x_2$
y2
Third stream detection
x3

Figure 3. Explanation of OSIC signal detection, an example of three transmitting antennas.

Otherwise, if the process of canceling the interference is done with an erroneous decision in any stage, error propagation will spread in the following stage by order of detection. Hence, next in this section, we describe two methods for reducing error propagation.

- SINR-Based Ordering (SINR-BO): in this case, the stream with big post detection Signal-to-Interference-plus-Noise-Ratio (SINR) is detected first [1]. Based on the transformation matrix $T_{MMSE} = (\widehat{\mathbb{H}}^H \widehat{\mathbb{H}} + 2\sigma_n^2 I)^{-1} \widehat{\mathbb{H}}^H$, the post detector with SINR is defined as:

$$SINR_i = \frac{E_x |T_{i,MMSE} \widehat{\mathcal{H}}_i^{UCA,XPD}|^2}{E_x \sum_{l \neq i} |T_{i,MMSE} \widehat{\mathcal{H}}_l^{UCA,XPD}|^2 + \sigma_n^2 ||T_{i,MMSE}||^2} \tag{29}$$

where, $i = 1,2, \cdots, N_t$ and E_x transmitted signal energy. $T_{i,MMSE}$ is the i^{th} row of T_{MMSE} and $\widehat{\mathcal{H}}_i^{UCA,XPD}$ is the i^{th} column vector of the estimated channel matrix $\widehat{\mathbb{H}}$. In fact, once the N_t of SINR are calculated based on T_{MMSE}, we extract the equivalent stage with the maximum SINR. In addition to that, the procedure discussed above is applied for symbol detection. Furthermore, T_{MMSE} is modified by suppression of the channel gain vector equivalent to the data detected. Otherwise, the computational complexity of all numbers of SINR is given by $\sum_{i=1}^{N_t} i = \frac{N_t(N_t+1)}{2}$.

- SNR-Based Ordering (SNR-BO): in this method, higher Signal-to-Noise-Ratio (SNR) is detected first [1]. Similarly based on the transformation matrix T_{ZF}, SNR is defined as:

$$SNR_i = \frac{E_x}{\sigma_n^2 ||T_{i,ZF}||^2} \tag{30}$$

where, $i = 1,2, \cdots, N_t$. Similarly, the procedure discussed in the first method can be used. Otherwise, the computational complexity of all numbers of SNR is giving by $\sum_{i=1}^{N_t} i = \frac{N_t(N_t+1)}{2}$ [1].

6. Simulation Results

In the next part of this paper, a set of performance results is discussed. According to Figure 1, the length of OFDM subcarriers is equal to 512 and the CP is 128; the higher order modulation QAM is taken equal to 64, the g consecutive OFDM symbol is set to 100 and the number of taps is supposed equal to 1. Furthermore, based on LSCE method (Section 3), the estimated channel is evaluated. Our proposed UCA-mMIMO system is analyzed for 10000 samples of the channel based on Monte Carlo simulation and the power is normalized. Moreover, Figure 4 depicts channel orthogonality (i.e., favorable propagation) for the UCA-mMIMO system on the one hand. On the other hand, the XPD is set to be 8 dB and the mean of AAoA/EAoA is equal to 0Â°; the distance between the terminal and the BS antenna is set to be 200λ. Similarly to some reports in the literature [5], [33]-[35], the antenna spacing is equal to 0.5λ.

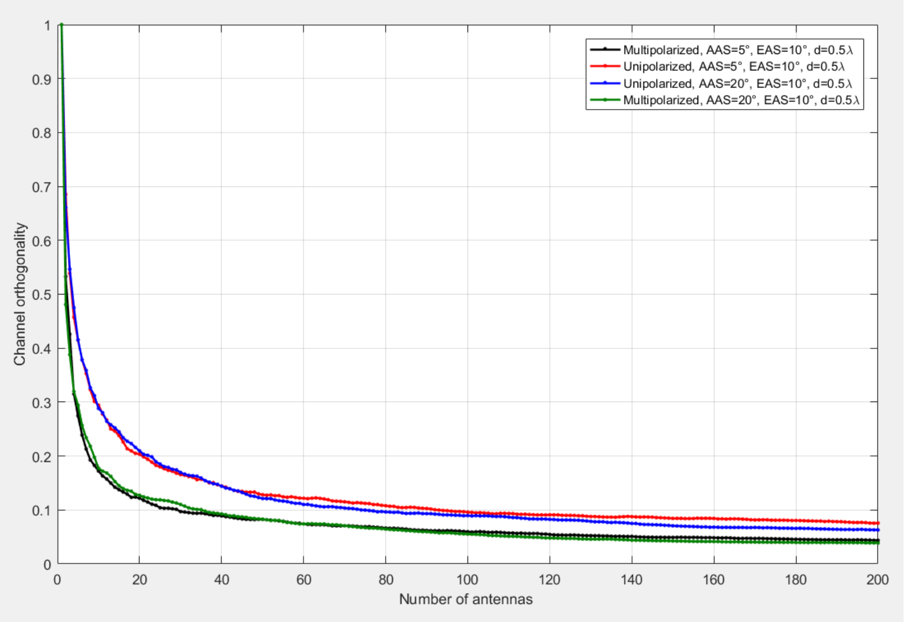


Figure 4. Channel orthogonality *vs.* number of antennas with different AAS.

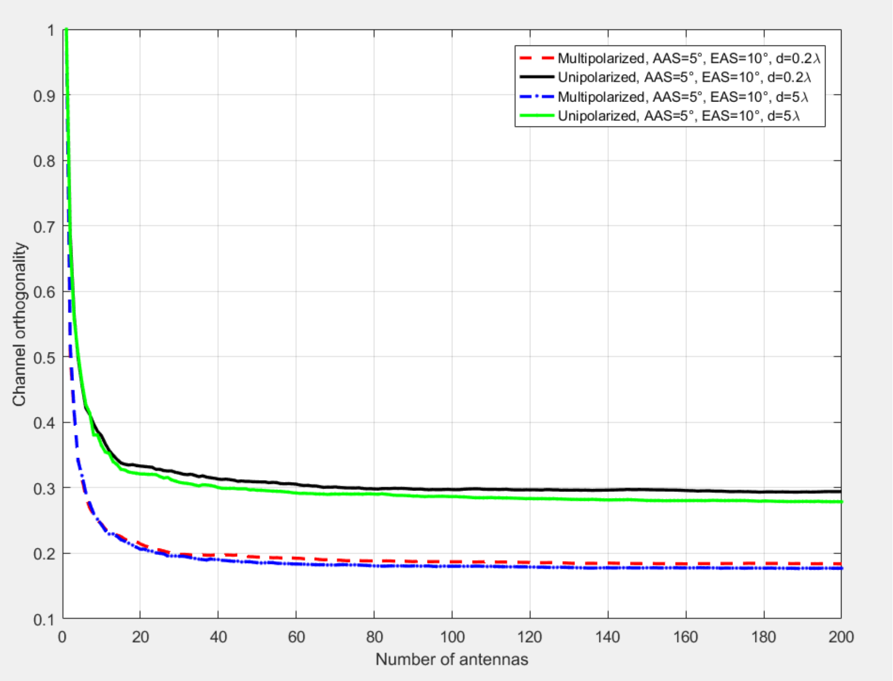


Figure 5. Channel orthogonality *vs.* number of antennas with different antenna spacings.

From Figure 4, it has been found that channel orthogonality of multipolarized/unipolarized systems decreases when the number of antennas increases. In the case when AAS=5Â° (i.e., poor scattering) and $N_r = 20$, channel orthogonality is nearly 0.2 for unipolarized systems; while it's nearly 0.12 for multipolarized systems; in the same way, when AAS=20Â° (i.e., rich scattering) and $N_r = 200$, channel orthogonality is nearly 0.06 for unipolarized systems, while it's nearly 0.03 for multipolarized ones. Consequently, channel orthogonality is affected by AAS, while the effect of EAS is negligible. In addition to that, using multipolarized UCA-mMIMO can decrease more channel orthogonality (i.e., favorable propagation) compared to unipolarized UCA-mMIMO. Hence, employing multipolarized UCA-mMIMO system in real environment can decline the necessity of rich scattering.

In the next part of this paper, AAS is equal to $5\hat{A}°$ and EAS is equal to $10\hat{A}°$. Also, XPD is set to be 8 dB. Figure 5 shows that channel orthogonality declines as the BS antenna spacing (i.e., N_r) increases. In the case with $N_r = 20$ and $d = 0.2\lambda$ (i.e., small antenna spacing), channel orthogonality is nearly 0.33 with unipolarized antennas, while it's nearly 0.21 for multipolarized antennas. Moreover, when $N_r = 200$, channel orthogonality is nearly 0.29 with unipolarized antennas and it's about 0.18 with multipolarized antennas. Otherwise, when $d = 5\lambda$ (i.e., large antenna spacing) the multipolarized antennas decline more channel orthogonality compared to unipolarized antennas. Hence, using small antenna spacing and multipolarized UCA-mMIMO can help decrease channel orthogonality between terminals and reduce the need of a large antenna spacing.

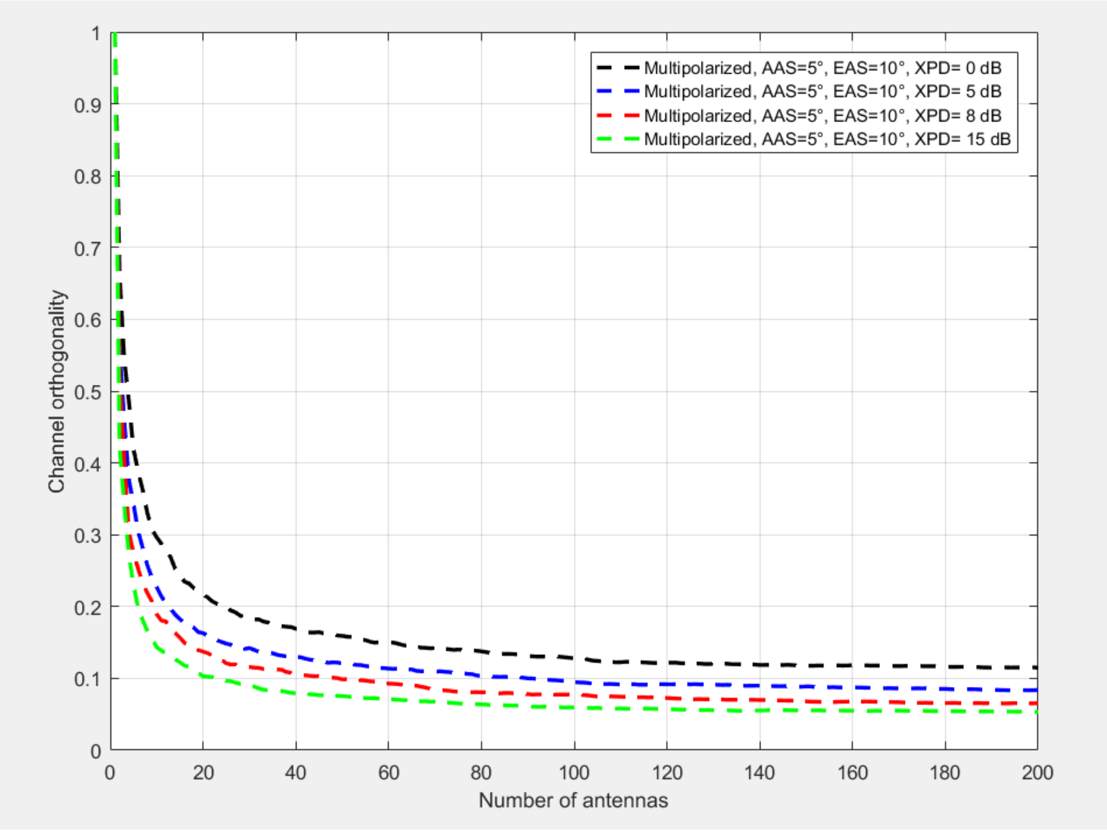


Figure 6. Channel orthogonality *vs.* number of antennas with different XPD.

Otherwise, in the case of poor scattering (i.e., AAS=5Â°) and an antenna spacing kept at the same value (0.5λ), the Figure 6 shows channel orthogonality of multipolarized UCA-mMIMO system; it has been found that when $N_r = 20$, channel orthogonality is nearly 0.22 with $XPD = 0$ dB and it's nearly 0.1 with $XPD = 15$ dB. When the BS antenna $N_r = 200$, channel orthogonality with $XPD = 0$ is nearly 0.12, while it's lower than 0.05 with $XPD = 15$ dB. Furthermore, an increase of XPD declines more channel orthogonality, thus more power is kept in the similar transmitting and receiving polarization states. Similarly, a large BS antenna number can help decrease channel orthogonality, specifically in a small XPD (i.e., more power seepage).

Figure 7 shows the performance results of OSIC detector using multipolarized/unipolarized UCA-mMIMO system. AAS is equal to 5Â° and EAS is equal to 10Â°; XPD is set to 8 dB. In addition to that, $d = 0.5\lambda$ and $d_y = 200\lambda$; according to estimated matrix (25), the number of terminals (N_t) is equal to 50 and the BS antenna number (N_r) is equal to 200. From this Figure the BER decreases over the range of SNR. Furthermore, in high-SNR region, multipolarized UCA-mMIMO systems perform better than unipolarized UCA-mMIMO systems; at SNR equal to 20 dB and using $OSIC\_SINR$ based ordering, BER is equal to 21.78×10^{-3} and 20.55×10^{-4} for unipolarized and multipolarized antennas, respectively, while when $OSIC\_SNR$ based ordering is used, BER is equal to 78.98×10^{-4} and 70.94×10^{-5} for unipolarized and multipolarized antennas, respectively. Hence, multipolarized antennas outperform unipolarized antennas; in addition to that, the gaps between the true channel and multipolarized antennas using $OSIC\_SINR$ based ordering and $OSIC\_SNR$ based ordering are equal to 2 dB and 0.8 dB, respectively at $BER = 18.7 \times 10^{-4}$. Consequently, $OSIC\_SNR$ based ordering with multipolarized UCA-mMIMO system provides a better performance.

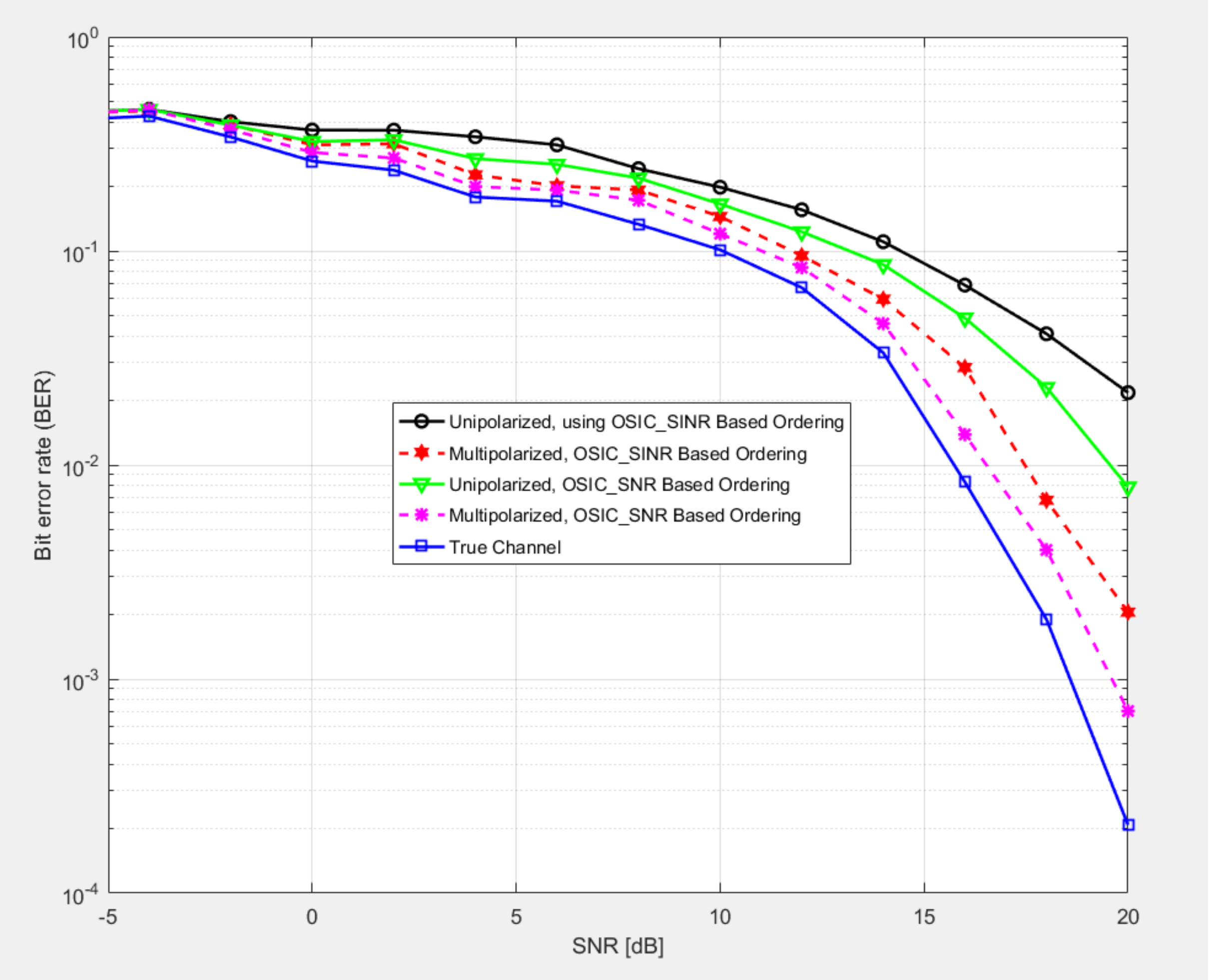


Figure 7. Bit error rate *vs.* signal to noise ratio for unipolarized/multipolarized UCA-mMIMO system with two OSIC detectors.

7. Conclusions

In this paper, UCA-mMIMO system design is proposed. A 3-D channel pattern of UCA-mMIMO is estimated using LSCE method. The presented pattern can be adjusted based on parameters analyzed. From the simulation results, it can be concluded that multipolarized UCA-mMIMO in each time declines more channel orthogonality in many situations compared to unipolarized UCA-mMIMO. Using $OSIC\_SNR$ based ordering with multipolarized UCA-mMIMO system provided a better performance compared to $OSIC\_SINR/OSIC\_SNR$ based ordering with unipolarized UCA-mMIMO system. Summing up the results, it can be concluded that our proposed pattern using multipolarized antennas can be implemented and adapted if miniaturization of electronic elements is indispensable and would be the first candidate for Massive-MIMO systems.

REFERENCES

[1] A. Riadi, M. Boulouird and M. M. Hassani, "ZF/MMSE and OSIC Detectors for UpLink OFDM Massive MIMO systems," Proc. of the IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), pp. 767 – 772, 9-11 April 2019, Amman, Jordan, 2019.

[2] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfrojd and T. Svensson, "The Role of Small Cells, Coordinated Multipoint and Massive MIMO in 5G," IEEE Communications Magazine, vol. 52, no. 5, pp. 44-51, 2014.

[3] E. Bjornson, E. G. Larsson and T. L. Marzetta, "Massive MIMO: Ten Myths and One Critical Question," IEEE Communications Magazine, vol. 54, no. 2, pp. 114-123, 2016.

[4] L. Zhao, K. Li, K. Zheng and M. Omair Ahmad, "An Analysis of the Tradeoff between the Energy and Spectrum Efficiencies in an Uplink Massive MIMO-OFDM System," IEEE Transactions on Circuits and Systems, vol. 62, no. 3, pp. 291-295, 2014.

[5] F. Rusek, D. Persson, B. Kiong Lau, E. G. Larsson, T. L. Marzetta, O. Edfors and F. Tufvesson, "Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays," IEEE Signal Processing Magazine, vol. 30, no. 1, pp. 40-60, January 2013.

[6] T. Van Chien and E. Björnson, "5G Mobile Communications," *Springer*, pp. 77-116, 2017.

[7] H. Quoc Ngo, E. G. Larsson, T. L. Marzetta and M. Omair Ahmad, "Energy and Spectral Efficiency of Very Large Multiuser MIMO Systems," IEEE Transactions on Communications, vol. 61, no. 4, pp. 1436 - 1449, 2013.

[8] A. Riadi, M. Boulouird and M M.Hassani, "Least Squares Channel Estimation of an OFDM Massive MIMO System for 5G Wireless Communications," in: M. Bouhlel, S. Rovetta (Eds.), Proceedings of the 8th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT'18), vol. 2, 2018, Smart Innovation, Systems and Technologies, vol. 146, Springer, Cham.

[9] S. Yang and L. Hanzo, "Fifty Years of MIMO Detection: The Road to Large-Scale MIMOs," IEEE Communications Surveys & Tutorials, vol. 17, no. 4, 2015.

[10] A. Riadi, M. Boulouird and M. M. Hassani, "An Overview of Massive-MIMO in 5G Wireless Communications," Colloque International TELECOM 2017 & 10 emes JFMMA, EMI - Rabat, Morocco, Mai 10-12, 2017.

[11] P. Rajeev, A. Prabhat and L. Norsang, "Sphere Detection Technique: An Optimum Detection Scheme for MIMO System," International Journal of Computer Applications, vol. 100, no. 2, pp. 975-987, August 2014.

[12] D. Wiibben, R. Bohnke, V. Kuhn and K.-D. Kammeyer, "MMSE Extension of V-BLAST Based on Sorted QR Decomposition," Proc. of IEEE 58th VTC-Fall, pp. 508-512, Orlando, FL, USA, Oct. 2003.

[13] R. Bohnke, D. Wubben, V. Kuhn, and K.-D. Kammeyer, "Reduced Complexity MMSE Detection for BLAST Architectures," Proc. of IEEE GLOBECOM, pp. 2258-2262, San Francisco, USA, Dec. 2003.

[14] X. Cheng, Y. He, Li Zhang and J. Qiao, "Channel Modeling and Analysis for Multipolarized Massive-MIMO Systems," International Journal of Communication Systems, vol. 21, no. 12, pp. e3703, 2018.

[15] X. Cheng and Y. He, "Channel Modeling and Analysis of ULA Massive-MIMO Systems," Proc. of the 20th International Conference on Advanced Communication Technology (ICACT), pp. 411-416, 2018.

[16] A. Moradi, H. Bakhshi and V. Najafpoor, "Pilot Placement for Time-Varying MIMO OFDM Channels with Virtual Subcarriers," Communications and Networks, vol. 3, no. 1, pp. 31-38, February 2011.

[17] I. Barhumi, G. Leus and M. Moonen, "Optimal Training Design for MIMO OFDM Systems in Mobile Wireless Channels," IEEE Signal Processing Magazine, vol. 51, no. 6, pp. 1615-1624, May 2003.

[18] T.-L. Tung, K. Yao and R. E. Hudson, "Channel Estimation and Adaptive Power Allocation for Performance Arid Capacity Improvement of Multiple-Antenna OFDM Systems," Proc. of the 3rd IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC'01), China, 2001.

[19] A. Youssefi and J. El-Abbadi, "Pilot-symbol Patterns for MIMO OFDM Systems under Time Varying Channels," International Conference on Electrical and Information Technologies (ICEIT), Marrakech, Morocco, March 25-27, 2015.

[20] Y. S. Cho, J. Kim, W. Y. Yang and C. G. Kang, "MIMO-OFDM Wireless Communication Technology with MATLAB," Beijing: Publishing House of Electronics Industry, 2013.

[21] Y. He, X. Cheng and G. L. Stuber, "On Polarization Channel Modeling," IEEE Wireless Communications, vol. 23, no. 1, pp. 80-86, February 2016.

[22] M. Coldrey, "Modeling and Capacity of Polarized MIMO Channels," Proc. of VTC Spring 2008-IEEE Vehicular Technology Conference, pp. 440-444, May 2008.

[23] A. Habib, B. Krasniqi and M. Rupp, "Convex Optimization for Receive Antenna Selection in Multi-Polarized MIMO Transmissions," Proc. IEEE Sys., Signals and Image Processing Conf., pp. 269-275, April 2012.

[24] K. H. Jeon, B. Hui, K. H. Chang, H. S. Park and Y. O. Park, "SISO Polarized Flat Fading Channel Modeling for Dual-Polarized Antenna Systems," The International Conference on Information Network, pp. 368-373, Feb. 2012.

[25] L. Jiang, L. Thiele and V. Jungnickel, "On the Modelling of Polarized MIMO Channel," Proc. Europ. Wireless, pp. 1-4, 2007.

[26] N. H. Quoc, E. G. Larsson and T. L. Marzetta, "Aspects of Favorable Propagation in Massive-MIMO," Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), pp.76-80, September 2014.

[27] K. Zheng, S. Ou and X. Yin, "Massive-MIMO Channel Models: A Survey," International Journal of Antennas and Propagation, vol. 11, pp. 1-10, 2014.

[28] J. Park and B. Clerckx, "Multi-user Linear Precoding for Multi-polarized Massive- MIMO System under Imperfect CSIT," IEEE Transactions on Wireless Communications, vol. 14, no. 5, pp. 2532-2547, 2015.

[29] A. Chockalingam and B. Sundar Rajan, Large MIMO Systems, Cambridge University Press, 2014.

[30] G. J. Foschini, "Layered Space-time Architecture for Wireless Communication in a Fading Environment When Using Multi-element Antennas," Bell Labs Technical Journal, vol. 1, no. 2, pp. 41-59, 1996.

[31] P. W. Wolniansky, G. J. Foschini, G. D. Golden and R. A. Valenzuela, "V- BLAST: An Architecture for Realizing Very High Data Rates over the Rich-scattering Wireless Channel," Proc. of URSI International Symposium on Signals, Systems and Electronics Conference Proceedings (Cat. No.98EX167), pp. 295-300, 1998.

[32] G. D. Golden, C. J. Foschini, R. A. Valenzuela and P. W. Wolniansky, "Detection Algorithm and Initial Laboratory Results Using V-BLAST Space-time Communication Architecture," Electronics Letters, vol. 35, no. 1, pp. 14-16, 1999.

[33] J. Li, Y. Zhao and Z. Tan, "Indoor Channel Measurements and Analysis of a Large-scale Antenna System at 5.6 GHz," Proc. of IEEE/CIC International Conference on Communications in China (ICCC), pp. 281-285, 2014.

[34] X. Gao, O. Edfors, F. Rusek and F. Tufvesson, "Linear Pre-coding Performance in Measured Very-Large MIMO Channels, Proc. of IEEE Vehicular Technology Conference (VTC Fall), pp.1-5, 2011.

[35] J. Hoydis, C. Hoek, T. Wild and S. Ten Brink, "Channel Measurements for Large Antenna Arrays," Proc. of International Symposium on Wireless Communication Systems (ISWCS), pp. 811-815, 2012.

ملخص البحث:

في هذه الورقة، يتم اقتراح تصميم مبتكر لنظام ضخم ذي مصفوفة دائرية متجانسة، متعدد المداخل والمخارج، مبني على الموجة الكروية، في الإرسال القائم على الربط العلوي؛ إذ يتم إنشاء نمط ثلاثي الأبعاد للقنوات، وتحليل تعامُد القنوات في الأنظمة متعددة الاستقطاب والأنظمة أحادية الاستقطاب متعددة المداخل والمخارج، التي تمتلك مصفوقة دائرية متجانسة. فقد جرى تقييم أنظمة متعددة الاستقطاب وأخرى أحادية الاستقطاب؛ من أجل التقليل من تعامُد القنوات.

وقد تم أخذ كل من: زاوية سَمْت الوصول، وزاوية مَيْل الوصول، جنباً الى جنب مع مسافة التباعُد بين الهوائيات، وتمييز الاستقطاب المتقاطع، بعين الاعتبار. وباستخدام طريقة مونتِ كارلو للمحاكاة، بينت النتائج أن الأنظمة متعددة الاستقطاب تعطي أداء أفضل إذا قورنت بمثيلاتها من الأنظمة أحادية الاستقطاب في ظل أوضاع مختلفة.

والجدير بالذكر أن التصميم المقترح في هذه الدراسة يتعين تحقيقه بشكل بسيط في بيئة حقيقية تتلاءم مع المتغيرات التي خضعت للتحليل؛ من أجل التحقق من أنه يمثل خياراً جيداً جداً.

AN ENERGY-EFFICIENT QUALITY OF SERVICE (QOS) PARAMETER-BASED VOID AVOIDANCE ROUTING TECHNIQUE FOR UNDERWATER SENSOR NETWORKS

Kamal Kumar Gola[1] and Bhumika Gupta[2]

(Received: 12-Jul.-2019, Revised: 16-Sep.-2019 and 21-Sep.-2019, Accepted: 8-Oct.-2019)

ABSTRACT

Underwater sensor networks (UWSNs) have become among the most interesting research areas, since they open the door wide to researchers to conduct research in this field. There are so many issues in underwater sensor networks. The most serious issue is the void region that degrades the performance of networks. It is an issue, where a node doesn't have any forwarder node to forward the packets to another node. Here, the objective of this work is to avoid the void region. For the same purpose, this work proposes an algorithm named "An Energy-Efficient Quality of Service (QoS) Based Void Avoidance Routing Technique". The proposed work uses two-hop node information to avoid the problem of void region. This approach uses depth information, distance to next, holding time and residual energy as Quality of Service (QoS) parameters in order to find the best forwarder node to forward the data packets to their destination. The proposed algorithm has been implemented in MATLAB. Results show a better performance in terms of packet delivery ratio, energy tax and number of dead nodes as compared to Energy-Efficient Void Avoidance Routing Scheme for Underwater Wireless Sensor Network (E2RV).

KEYWORDS

Void avoidance, Sensors, Coverage, Connectivity, Surface station, Holding time, Quality of Service (QoS) parameters.

1. INTRODUCTION

Underwater Sensor Networks (UWSNs) are regarded among the favourable technologies for accumulating beneficial and valuable data from underwater seas. This technology mainly helps in assisting environmental predictions and military operations and comprises of underwater sensor nodes. One of the most challenging issues in underwater sensor networks is considering the demand and need to forward data packets with high packet delivery ratio and minimal energy consumption [1]-[2]. These sensor nodes are placed at water surface and at different depths of water. In Underwater Sensor Networks (UWSNs), the challenging task is to send data timely and efficiently because of the complex underwater environment. Radio signals are not fruitful in underwater environment because of radio signals' rapid attenuation in underwater environment [3]. So, acoustic signals are used for underwater communication. When data reaches a sink, then radio waves are used to forward data packets to the base station. In underwater environment, water current sensor nodes can move and change their position, which leads to a dynamic network topology. Apart from this, the links between sensor nodes are highly error prone due to path loss. There are some challenges related to acoustic communication as compared to electromagnetic waves, like: path loss, high error rate, propagation delay, lower bandwidth …etc. [4]-[5]. In acoustic communication, path loss can be defined by distance and frequency as:

$$P\_Loss\,(d, f) = P\ Loss_0\ d^k\ a\ (f) \tag{1}$$

where, d is the distance, f indicates the signal frequency, k is known as the spreading factor and the value of the spreading factor depends on the spherical factors and practical spreading. $PLoss_0$ signifies a constant and a (f) is the absorption coefficient as defined in [6]. The formula below is valid for high frequency.

$$10log_a(f) = 0.11f^2/(1+f^2) + 44 * f2/(4100 + f2) + 2.75 * 10^{-4}f^2 + 0.003 \tag{2}$$

[1] K. K. Gola is with Department of Computer Science and Engineering, Uttarakhand Technical University, Dehradun and Faculty of Engineering, TMU, Moradabad, India. Email: kkgolaa1503@gmail.com

[2] B. Gupta is with Department of Computer Science and Engineering, G.B. Pant Institute of Engineering & Technology, Uttarakhand, Pauri Garhwal, India. Email: mail2bhumikagupta@gmail.com

For low frequencies, the formula below is suitable.

$$10log_a(f) = 0.11f^2/(1+f^2) + 0.011 * f^2 + 0.002 \qquad (3)$$

In underwater communication, there are two types of noise; natural noise which can be generated by fish and tide rain and man-made noise that can be generated by ship movement. Delay is also a major challenge in underwater communication. Generally, the speed of acoustic waves is 1500m/s having a delay of 0.67s/km. Some parameters, like pressure, temperature and depth, affect the sound velocity in underwater communication. The formula below is used to calculate the velocity if pressure and temperature is known.

$$C = 1449.2 + 4.6T - 0.055T^2 + 0.00029T^3 + (1.34 - 0.010T)(S - 35) + 0.016d \qquad (4)$$

Here, the temperature (T) is in centigrade (°C), S represents the salt in water in parts per thousand (‰), d indicates the depth which is calculated in meters and c is the sound velocity which is given in meter per second. The formula above is suitable for the following conditions [7]:

Routing depends on the selection of next forwarding node, which is the key component of Underwater Sensor Networks (UWSNs) having a direct and dominating effect on packet delivery ratio and energy consumption. Consequently, to resolve this problem, the research community decided to enhance the performance of Underwater Sensor Networks (UWSNs) [8]-[9]. Apart from this, underwater sensor networks also include some major challenges that are mentioned below:

- Sensor node deployment.
- Network connectivity.
- Limited energy.
- Low data rate due to acoustic communication.
- Large propagation delay.
- Unpredictable underwater environment.
- High equipment deployment cost.
- Unscalability.
- No interaction.
- Complicated design and network deployment.

1.1 Problems Related to Selection of Next Forwarding Node

The major issue in routing technique is the selection of next forwarding node [10]-[12]. This matter captivated researches to generate next forwarding node algorithms. Firstly, each of the source nodes selects a group of its neighbours. The selection of forwarder node is based on different parameters, such as physical distance, residual length and link quality. These forwarder nodes can hold data packets for some predefined time based on some different parameters, like: sound propagation speed, range of transmission, distance to elude redundant packet transmission and collision. The node having the least holding time will be selected as the best forwarder node which forwards data packets to another node or to the sink node. The removal of data packets from their buffers takes place if and only if a sensor node overhears the transmitted data packet. Otherwise, it waits till its holding time terminates to forward the data packet. Consequently, the aforementioned algorithms have a direct influence on energy consumption and packet delivery ratio.

The use of different parameters for the selection of next forwarding nodes results in a direct influence on the entire performance of routing protocols. The energy between the nodes is balanced by the use of residual energy metric. Another major metric that has a direct effect on upgrading the packet delivery ratio and minimizing energy consumption is link quality. The usage of depth metrics minimizes the consumption of energy, because the calculation of each node takes place locally. Physical distance can be measured using the beaconing messages assigned by the sink. Due to impotence of Global Positioning System (GPS) in underwater circumstances, identification of node location is costly using Global Positioning System (GPS). Consequently, it becomes important to derive and design an algorithm that selects the forward node based on multi-metrics, specifying energy-efficient and authentic forwarding nodes in order to minimize the consumption of energy, reduce traffic on the designated network and ensure the proper delivery of data packets [13]-[14].

One of the most major issues that came up on next forwarding node selection is communication void. The next problem arises due to the absence of neighbour nodes in the transmission range [15]-[16]. Since Underwater Sensor Networks (UWSNs) sustain from a sparse and dynamic network structure, such structure obtrudes a low packet delivery ratio, further reducing throughput because of shortfall of algorithms that look for the communication gap. Furthermore, the manoeuvre of weak void avoiding algorithms elicits high dissipation of energy. Terrestrial wireless sensor networks (TWSNs) do not support void handling algorithms in harsh conditions due to some precise features of UWSNs. Hence, avoidance of void nodes in the operation of selection of forwarding nodes is significant in order to refine and enhance the packet delivery ratio and fortify the delivery of data packets.

The next significant issue involves the selection of the shortest path and has a direct effect on transmission number, network lifetime and energy consumption. For minimizing retransmission and number of nodes stipulated to transfer data, selection of the node having a lower depth as compared to sender is a conventional method to handle this issue. The foremost shortcoming of this algorithm is that it completely relies on distance without considering any other parameters, such as link quality and residual energy. Therefore, it becomes significant to ensure path selection while choosing the next forwarding node. This results in minimization of total forwarded packets, vanquishing the irrelevant forwarding and compressing the total dissipation of energy.

1.2 Underwater Sensor Network Architectures

1.2.1 1-D Architecture

In one-dimensional architecture, sensor nodes are individually deployed and equally responsible to sense data, process data and finally transmit data packets to the sink node or to the base station. For a particular time, the sensor node may float in underwater environment to sense the phenomena and then float again towards the upper direction to forward the sense data to the sink node or to the base station. Here, sensor nodes may communicate *via* radio frequency, acoustic or optical communication link [17].

1.2.2 2-D Architecture

In this architecture, nodes are placed in cluster form. Each cluster has one cluster head. Each cluster member senses the underwater phenomena and forwards the information to the cluster head. Now, the cluster head forwards information to surface buoyant nodes. Two types of communication take place in this architecture: horizontal communication link is used between cluster members and cluster head, while vertical communication link is used between cluster head and surface buoyant node. Nodes which are deployed in the depth can communicate with each other using acoustic communication and nodes which are placed at water surface can use radio communication. This type of architecture can be suitable for time-critical and delay-tolerant applications [18].

1.2.3 3-D Architecture

In this architecture, all sensor nodes are placed in cluster form and are anchored at different depth levels. Here, there are three types of communication scenarios that take place: first is intercluster communication among nodes, second is intracluster communication that means communication between sensor node and anchor node and third is communication between anchor nodes and buoyant nodes. Acoustic, radio frequency and optical links can be used by any type of communication.

1.4.4 4-D Architecture

It is a combination of mobile underwater sensor networks and 3-D underwater sensor networks (UWSNs). This architecture uses Remotely Operative Underwater Vehicles (ROVs). All anchor nodes send data packets to the Remotely Operative Underwater Vehicles (ROVs) and ROVs forward this data to the base station. It may be any vehicle, ship, robot, …etc. Each sensor node is able to send data to the Remotely Operative Underwater Vehicle (ROV) directly according to how close a sensor node is to the Remotely Operative Underwater Vehicle (ROV). The communication between Remotely Operative Underwater Vehicles (ROVs) and sensor nodes depends on the data and the distance between data. If a sensor node is near to the Remotely Operative Underwater Vehicle (ROV) node and has large data, then the sensor node can use radio link, while if the distance is far and the sensor node has small data, then the sensor node can use acoustic link [19]-[20].

1.3 Underwater Sensor Network Applications

Underwater Sensor Networks (UWSNs) cover many applications, like: water quality, water temperature, monitoring of underwater pipelines, environment, mine detection, disaster prevention, security, …etc.

Monitoring Applications: these applications refer to monitoring related to the environment, properties, characteristics and objects of interest under water. These applications are particularly related to monitoring of physical environment. These are further divided into three applications. First is monitoring **water quality:** water is one of the most valuable resources that are primary requirements for living under water and above water. So, it is necessary to monitor water. Second is monitoring the **habitat:** this application deals with the environment of living organisms. If considered under water, then it becomes more challenging due to underwater environment. Habitat monitoring if further divided into **reef**, **marine life** and **fish farm** monitoring. Third is monitoring the **underwater exploration:** this application refers to monitoring minerals available under water, like oil and gas. It is required to monitor underwater pipelines, because these pipelines are used for oil and gas. This application is further divided into **natural resource** and **pipeline and cable** monitoring. **Disaster Applications:** these applications are very critical, because they are very dangerous and may produce destruction on earth. Generally, disasters are unavoidable. This monitoring is related to events that aggravate water. Monitoring of **floods, volcanoes, earthquakes and tsunamis** comes under these applications. **Military Applications: t**hese are major applications of underwater sensor networks, where nodes are deployed to detect different features of military-related activities. Generally, Autonomous Underwater Vehicles (AUVs) are used to find submarines, secure ports, mines, …etc. Some sensing devices, like metal detectors and cameras, are equipped with AUVs for surveillance purposes. So, there are many potential scenarios, where (UWSNs) can be used. Here, Table 1. shows some major applications of underwater sensor networks.

Table 1. Underwater sensor network applications.

Application	**Objectives**	**Category**
Monitoring environment and geographical processes	i) Collecting information related to water. characteristics like: salinity, oxygen level and temperature. ii) Geographical processes include: earthquakes, tsunamis, floods and volcanoes.	Scientific
Counting or imaging animal life	Counting or imaging mammals, fish and microorganisms.	Scientific
Mine detection	Finding the location or position of specific mines in underwater environment.	Industrial, Military and Security
Minerals and oil extraction	First, finding the sources of minerals and oil and then installing the equipment in underwater environment to extract them.	Industrial
Monitoring underwater pipelines	Monitoring the underwater pipelines and providing immediate repairs.	Industrial
Offshore exploration	This application includes the exploration of unexpected regions.	Military and Security
Navigation	Providing accurate navigation to battle ships.	Military and Security
Communication	Providing communication between drivers and submarines.	Military and Security

2. PREVIOUS WORK

In [21], the authors have proposed an approach to overcome the problem of recovery and maintenance of the routing path. As a much less number of sensor nodes is needed to forward data, there is no need of sensor node state information. An interleaved and redundant path is used to send data from source to sink to overcome the problem of node failure and packet loss. In this approach, each node knows its location and the location of each data packet consists of the locations of all the involved nodes, like: source node, intermediate nodes and sink node. In this approach, a vector pipe is used to forward data from source to destination. Nodes that are near to this pipe are able to send data from source to sink. This approach reduces the network traffic and manages the topology. Like other approaches, this approaches has some drawbacks like that a virtual pipe can disturb the routing efficiency. It may be possible that if nodes become sparser due to water movement, it may be possible that a less number of sensor nodes will lie near or within the pipe or it may be possible that some paths will lie far from the pipe. In this situation, much less or no nodes are near to the virtual pipe, which results in low data delivery or no data delivery. Here, some nodes are involved again and again to send data, which results in more battery power consumption. Due to having a three-way handshaking nature, this approach has more communication overhead.

In [22], the authors have proposed a routing protocol to overcome the problem of continuous node movement. Sensor node uses dynamic addressing to resolve the problem of water currents. Due to this, the sensor node gets a new address based on a new position at a different depth level. According to this protocol, lots of sinks are fixed at water surface to collect data packets and some nodes are deployed at the bottom. Apart from this, the rest of nodes are deployed from bottom to surface at different-different levels. All nodes that are nearest to the surface have small addresses and become larger when the nodes float towards the bottom. This approach completes in two steps: first is to assign the dynamic addresses to the sensors node and second is to forward data using these dynamic addresses. These addresses are assigned by using hello packets that are generated by surface sinks. Each node tries to forward data packets to the upper direction in greedy fashion. The advantage of this protocol is that it does not require any special hardware or location information. Apart from this, it can handle the node movements. However, multi-hop routing problem still exists.

In [23], the authors have proposed Sector Based Routing with Destination Location Prediction (SBR-DLP) routing protocol. In this protocol, each sensor node knows its own location and guesses the destination node's location, where the precise information of the destination node's location is stored. All sensor nodes find their next hops using information received from the candidate nodes. This reduces the problem of multiple nodes that act as relay nodes. There is no need to rebroadcast the Request to Send (RTS). It is not possible to find the candidate node within the transmission range. Here, node speed causes disconnection, which results in low packet delivery ratio in sparse networks. If all nodes are mobile, then this protocol provides a better packet delivery ratio.

In [24], the authors have proposed an energy-efficient fitness-based routing protocol which uses depth, residual energy and distance from the forwarding node to sink. The proposed protocol does not use the control packet, which reduces energy consumption and end-to-end delay. A fitness function is also calculated to determine the best forwarder. The achievement of the proposed protocol is that it increases network lifetime and reduces end-to-end delay.

In [25], the authors have proposed a routing algorithm that uses residual energy and depth variance for void avoidance. The proposed approach uses the two-hope node information to remove the void holes in the network. It also uses depth and remaining energy of the node to transfer data from source to destination. On the basis of depth difference, this approach finds the best node to forward data packets. This approach also calculates the packet holding time in the network for a particular node. The proposed approach reduces the overall energy consumption and increases the network lifetime by distributing loads. The proposed technique shows less energy consumption and improves the network lifetime and packet delivery ratio. It also reduces data duplicity, but this approach has a high delay, which is the disadvantage of the proposed approach.

In [26], the authors have proposed a routing protocol that considers the knowledge of node localization and merges it with the network coding to reduce duplicate data packet transmission and energy consumption. The best forwarder is selected based on its location information, which reduces high energy consumption. To overcome the duplicate problem, the network coding is attached with the packets. Here, using an acknowledgement message, the end-to-end delay is high.

In [27], the authors have proposed a balanced energy-efficient routing protocol based on circular network. This circular network is divided into ten circular regions known as sectors. This protocol uses two mobile sinks to collect data from the sectors, which means that one mobile sink is used for five sectors. The mobile sink moves in a circular fashion to collect data from the nodes. This process balances energy consumption and enhances data packet reception by the sink node. No priority is assigned to the location by the sink node. Instead of this, a fixed pattern for the movement is followed. Due to this, packet loss and delay increase. This protocol is not suitable for sparse environments when nodes are deployed so far from each other.

In [28], the authors have proposed a protocol for maximum network coverage. The protocol uses two mobile sinks to collect data from the sensor nodes. The sinks move in a circular fashion in the network, which helps balance energy consumption and reduce packet loss. The disadvantage of the protocol is that as the sink node is moving in a circular fashion rather than in the targeted areas, there may be some nodes in the targeted areas that have to send data. Due to this, packet dropping increases and delay also increases, especially in sparse situations.

In [29], the authors have proposed a routing scheme for underwater wireless sensor networks. The main objective of this scheme is to find the best forwarder node to forward data packets to node/sink node/base station. This scheme uses the concept of Time of Arrival (ToA) and range-based equations to locate the sensor node in a recursive manner in the defined network. After localization, residual energy and coordinate of sensor nodes are used to find the best forwarder node. This scheme avoids horizontal transmission of data to reduce end-to-end delay. It also avoids the problem of void nodes and increases the network throughput. The 2-hop concept is used for better acknowledgement. The simulation results show good results in terms of energy consumption, Packet Delivery Ratio (PDR), average hop count, average end-to-end delay and propagation deviation factor.

In [30], the authors have proposed a routing scheme to select the best forwarder node by avoiding void node, or void region. The functionality of this scheme depends on two special parameters named: energy and depth. This scheme avoids horizontal transmission of data packets. The node having large residual energy and multiple neighbours will be selected as the best forwarder node. This scheme also calculates the holding time to reduce duplicate packet transmission. This scheme uses the concept of 2-hop neighbours. The results show that this scheme achieves 15% Packet Delivery Ratio (PDR) and propagates 40% less copies of data packets. The proposed scheme also reduces energy consumption, which increases the overall network lifetime.

In [31], the authors have proposed a location based routing protocol known as Power-Efficient Routing (PER). The primary objective of the protocol is to tackle the problem of energy consumption in underwater environment. Here, ordinary nodes are randomly scattered in the underwater environment, while the sink node is placed at the center. Sensor nodes that are available at the bottom are known as the source nodes. This protocol consists of two modules, where the first module is known as forwarder node selector and the other is known as forwarding tree trimming mechanism. This scheme uses the three parameters: residual energy, angle between two neighbours and distance to select the forwarder node. The selection of forwarder node is based on fuzzy logic technique. The second module is used when the number of duplicate packets is greater than the defined threshold. This protocol performs better as compared to Vector-Based Forwarding (VBF), as Power-Efficient Routing (PER) uses 2-hop nodes during the forwarding process, while Vector-Based Forwarding (VBF) uses the flooding technique in its virtual shape.

In [32], the authors have proposed a protocol to tackle the problem of communication void. Vector-Based Void Avoidance (VBVA) is another escalation of Vector-Based Forwarding (VBF) and is supposed to be the foremost void avoidance protocol that gives 3D flooding procedures to deal with the issue of communication void. Absence of void area possesses the same functionality of these protocols. The principle dissimilarity between these protocols is that Vector-Based Void Avoidance (VBVA) imparts two 3D flooding mechanisms to handle the communication void; known as vector-shift and back-pressure mechanisms. In vector shift, data packets are routed to the forwarder node available in the boundary of the void region. There are two conditions of being void. If void is convex, data packets can route towards the destination, while in case of concave void, the vector shift mechanism does not work efficiently. This scheme uses the back pressure mechanism to route back data packets in order to avoid the problem of concave void. This scheme achieves a good performance in mobile underwater sensor networks.

3. Problem Statement

The most challenging issue of routing protocols is communication void. The presence of void may cause packet delivery in the routing time, which leads to data loss. When the sender node does not trace any neighbour node in its range of transmission, this issue arises. This problem occurs when a sender node does not have any neighbour node or forwarder node to forward data packets towards the direction of sink node or surface station. Whenever this type of situation occurs, it is necessary to have an alternative path to forward data packets; otherwise, data packets will be discarded, which affects the network performance in terms of data loss.

4. Proposed Work

As a forementioned, nodes are originally positioned at the water surface and two-dimensionally connected network topology is formed and structured. Our elucidation fights vigorously to enlarge this two-dimensional network structure into the three-dimensional structure, so that the entire anticipated and sensed area of the three-dimensional network is increased and enhanced. Considering the elimination of coverage overlapping, the sensors that are nearer to each other are forwarded to distinct depths. The transmission of nearer sensors to distinctive depths facilitates the entire sensing coverage. Conversely, if the nodes are placed at a large distance to each other, the communication between them may be smashed and as a result, the ultimate network structure may possess internally disintegrated network segregation. Thence, it is of utmost significance to govern which conveyance linkage is to be safeguarded and which association and linkage may be shattered. Nevertheless, this issue is critical, conjecturing that the nodes are solely cognizant of the position of their one-hop proximate nodes. In our recommended algorithm, we assigned this significant charge to the surface station. Otherwise stated, the surface station will solely be accountable for the measurement of depth of every sensor in the mesh. The proposed work primarily encloses three important stages, which are:

1. Node Initialization.
2. Calculation of Depths of Nodes.
3. Distribution of Depths to Nodes.

4.1 Initialization Phase

This stage can be thought of as the starting setup stage. For the measurement of depths at the surface station, it is necessary to familiarize the surface station which the position of the sensors. When all the nodes are homogeneous and aimlessly dispersed at the surface of water in two dimensions, the sensors may not perceive the information regarding the location of the surface station. In this section or stage, a linked tree organization is devised, which is embedded at the surface station. The commencement of this stage takes place by transmitting a message known as Tree Request (TREQ). To answer this request, the one-hop neighbours dispatch a tree join report to their parent to form the connected tree and transmit a fresh Tree Request (TREQ) message to recognize the next hop neighbours. All the while, each of the nodes present in the network discovers its respective parent and children. After a linked tree structure is established, each node transmits its coordinates to their respective parents unassisted. The parent nodes then accumulate the coordinates received in the sub-tree in addition to their own coordinates. This accumulated list of coordinates is then forwarded to their parents. This activity and operation continue till the sink node acquires the coordinates of all sensors connected in the tree.

4.2 Depth Calculation

After the initialization phase, depth calculation method is invoked to calculate the depth for each node.

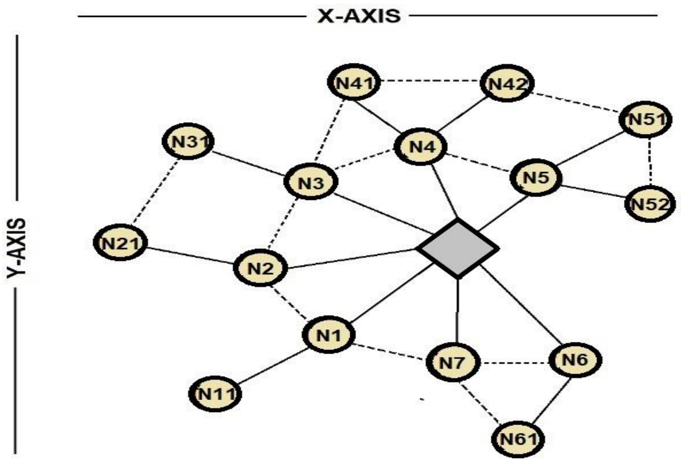


Figure 1. Nodes connected in 2-dimensional network topology.

The surface station is responsible for depth calculation till the first level. After this, the nodes at the first level are responsible for calculating depth for the next level. This process will continue till the last node of the list. Initially, all nodes are placed at the surface of water, where the z coordinate is assumed to be zero. Here, the surface station acts as a root and all other nodes will be left children or right children of the surface station. This process will make a connected tree structure and sensor nodes having highest depth will transfer their data to their upper nodes. Each upper node will act as a sub-root node for the same node. To calculate the sensor node depth, firstly the value of z coordinate for all sensor nodes is set to zero.

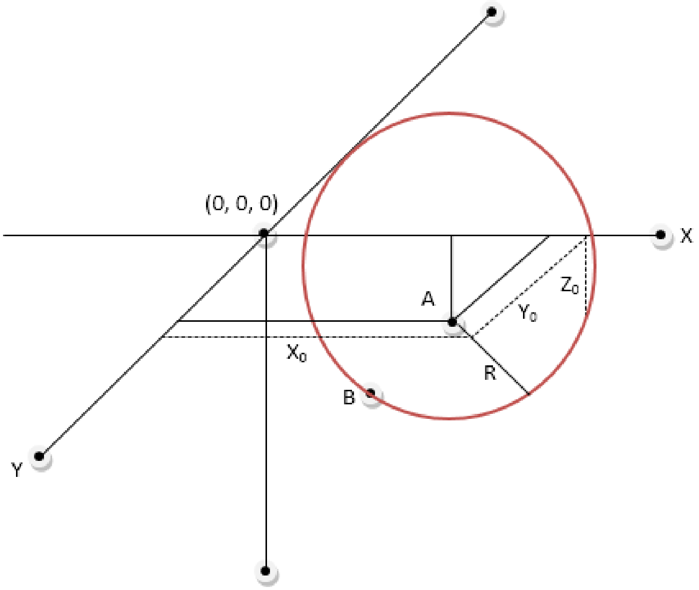


Figure 2. Deployment based on depth.

Here, R is the radius of the sphere and A is the center of the sphere. The coordinate of the surface station is A (x_0, y_0, z_0) and the coordinate of the sensor nodes is B (x, y, z). Initially, the value of z-coordinate for both surface station and sensor nodes is set to zero. The value of x- and y-coordinates for the sensor node B will remain the same when it will drift into the water, but the value of z-coordinate will change. The equations of the sphere are:

$$(x-x_0)^2+(y-y_0)^2+(z-z_0)^2=R^2 \tag{5}$$

$$(z-z_0)^2=R^2-(x-x_0)^2-(y-y_0)^2 \tag{6}$$

$$z=z_0+[R^2-(x-x_0)^2-(y-y_0)^2]^{1/2} \tag{7}$$

Using Equations 5 and 6, we can calculate the value of z-coordinate for the respective node, where the value of z-coordinate can be determined with the condition that the value of x- and y-coordinates of sensor node B is known. During calculation of depth for a particular node, we need to check the closest node and calculate the depth of that node relative to the closest node in order to reduce the problem of overlapping.

4.3 Depth Distribution to Nodes

Till all the nodes are processed, the customized depth evaluation stage goes on. After quantifying the depths of all the nodes, it is necessary to acknowledge the nodes regarding their evaluated depths. This phase requires the use of a linked tree by the surface station, which is devised after the initialization phase. Fundamentally, the surface station transfers the computed depths of sensor nodes to their neighbours in two-dimensionally. This message is encountered by the children of surface station, which then transfers the same message to their children. This procedure steadily goes on till all the nodes acquire the information concerning their depths. The nodes start submerging their specified depths to accomplish the deployment process as soon as the message is forwarded to their children.

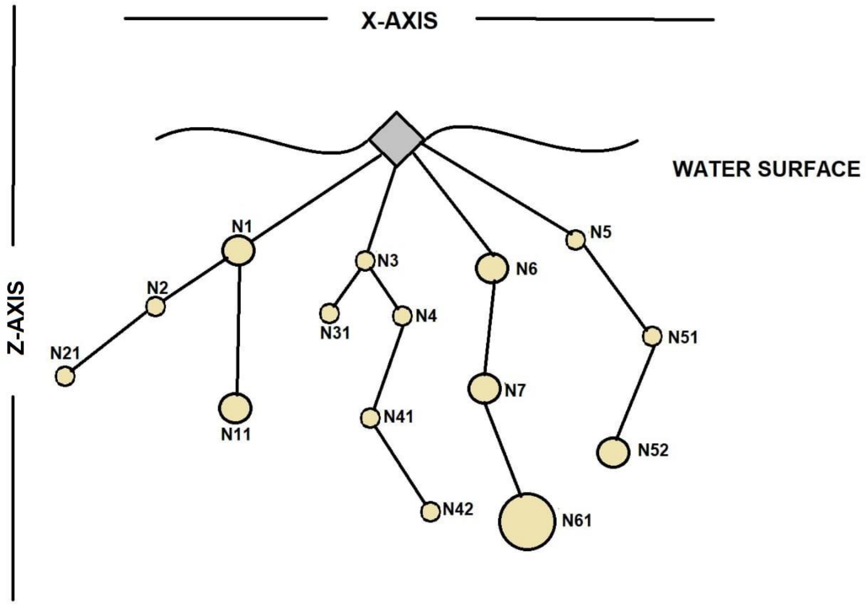


Figure 3. Nodes connected in 3-dimensional network topology.

To demonstrate the functionality of the algorithm, let us examine the example in Figure 1. Initially, sensor nodes are placed randomly at water surface and form a two-dimensionally-connected tree. Figure 1 shows that nodes are connected in two-dimensional form. As the surface station has completed the initialization phase, this results in the formation of a connected tree, as shown in Figure 1, having solid lines and the propinquity of nodes that instigates the overlapping problem in two dimensions is shown by dashed lines in Figure 1. Firstly, the surface station exercises N1 and calculates its depth using Equation (7). The depth of N1 respective to the surface station is represented in Figure 3. Then, the algorithm appends N1 in the processed list and processes N2. As N2 is near to N1 in 2-dimensional framework, to reduce coverage overlaps, these nodes need to be settled down at different depths. Hence, the algorithm calculates the depth of N2 correlative to N1 and puts it at a deeper level. As connectivity of N1 is ensured in the previous iteration, connecting N2 to N1 ensures its connectivity as well. Correspondingly, the algorithm links N2 to the process list. Back to back, the algorithm works on N3, which is near to N2 in 2-dimensional framework and calculates its respective depth to the station. N2 and N3 are positioned at distinct depths, which again reduces the overlapping problem between them. This algorithm calculates the depth of N4, N5, N6 and N7 in similar fashion. N11 is processed after processing the first-hop neighbors in 2-dimensional fashion. As N11 possesses a single neighbor in 2D, it is directly connected to N1. Equivalently, the computation of depth of N21 starts in the upcoming iteration by linking it with N2. Then, the algorithm works on N31, where two choices are generated: 1. Connecting N31 with N21; 2. Connecting N31 to N3. Since both of the choices discard the feasible coverage overlap, the algorithm puts N31 nearer to the surface and links it to node N3. The algorithm then works on N41, N42, N51, N52 and N61 in the same fashion.

4.4 Selection of Best Forwarder

As we know, the selection of best forwarder candidate is one of the most important tasks in any routing technique. It is assumed in the proposed approach that each node knows the location of the surface station. In our proposed approach, the selection of best forwarder is based on certain metrics like: depth variance, distance and holding time. Whenever a node needs to send data, firstly, the sender node computes the value of fitness function, adds the value of this function to its own location coordinate and then broadcasts the data packets. Only one-hop neighbours can receive this packet and compute their fitness function value with the sender one. If the calculated value of fitness function is greater than the sender's fitness function value that incorporates in the packet, it forwards the packet; otherwise, it discards it. There may be some situations where a more number of sensor nodes is involved to forward the data packet. To overcome this problem, this protocol calculates the holding time which is based on depth, residual energy and distance from the sender node to the forwarder node. The holding time may vary. The node having the less holding time forwards the packet, while the other nodes overhear the same data, avoiding the transmission.

Packet Format

A data packet consists of four fields as shown in Table 2.

Table 2. Packet format.

S_ID	P_SN	S_FV	S_L	Data

where, S_ID represents sender Id, P_SN is packet sequence number defined by sender node, S_FV is value of fitness function of sender and S_L represents the location of sender node.

Estimation of Fitness Function Value

A node having the greatest value of fitness function will be the best forwarder and the estimation of fitness function value is expressed in the following way:

$$f(n) = g(n) * h(n) \tag{8}$$

Calculation of g (n) Function

The calculation of g (n) function is expressed in the following way:

$$g(n) = Er(f) * depth\ (difference) * dsf \tag{9}$$

where, *Er(f)* represents the residual energy of the forwarding node, *depth(difference)* represents the difference of vertical differences from the forwarding node and the sending node to the surface and *dsf* is the distance from the sending node to the forwarding node.

Calculation of h (n) Function

The calculation of h (n) function is expressed in the following way:

$$h(n) = 1/\ dfd(node) \tag{10}$$

where, *dfd (node)* is the distance from the forwarding node to the destination node. Here, h (n) is estimated as the inverse of the distance from the forwarding node to the sink node, noting that less distance shows best node to forward the packet.

From the Equations 7 and 8, the fitness function stands in the following way:

$$f(n) = (Er(f) * depth(difference) * dsf)/\ dfd(node) \tag{11}$$

Estimation of Holding Time

To find the best forwarder, priority P will be assigned to these forwarders in order to forward data to the next forwarder. Remaining nodes that received the same packet will wait for a defined interval until they hear the same packet transmission from another node.

$$P = 1/\ g(n) \tag{12}$$

$$H_{Time} = [(1 - ((Er(f)/E))] * T_{Delay} * P \tag{13}$$

Here, H_{Time} = receiver's holding time, $Er(f)$ = residual energy, E = initial energy and T_{Delay} = predefined maximum delay.

At the point when a sensor node gets the data packet, it will look for the depth which is specified in the header of the data packet and then compare the depths. If it is situated at a higher depth, it will essentially discard the data packet. If it is located at a lower depth, it will calculate the distance from the source by using the concept of Time of Arrival (ToA). Distance between two nodes can be calculated by signal transmission time. Suppose t1 time is needed to send the packet and it reaches the destination at time t2 with velocity V, then the distance can be calculated as:

$$Dist\ to\ Next = V(t2 - t1) \tag{14}$$

Algorithm

```
Source node is ready to send the data packet
Receiver node receives the data packet
If $R_{(Depth)} > S_{(Depth)}$
Where $R_{(Depth)}$: Depth of receiving node and $S_{(Depth)}$: Depth of source node
Drop the data packet
else
Calculate the value of fitness function, set the priority and $H_{Time}$ (Holding Time) for data packet
end if else
Extract the value of fitness function and calculate own fitness function value
If calculated value>extracted value then
Receive the data packet
else discard the packet
end if else
while($H_{Time}$ expires)
listen the channel
if same packet overhear then
discard the data packet
else
forward the packet
end if
end while
```

4.5 Data Transmission

The proposed algorithm is designed to avoid horizontal transmission, which increases network lifetime and energy efficiency. Data transmission phase starts when a node has a data packet to send. The main objective is that the data packet should reach one of the sinks in multi-hop fashion. In the proposed approach, the number of acknowledgements per data packet may vary from zero to two hops based on the sender's and receiver's status. Here, zero-hop acknowledgement is used to control message transmission and end-to-end delay, while two-hop acknowledgement is used to handle the void node issue and improve the data delivery ratio. When a data packet is stuck in the communication void region, a control message-like acknowledgement is used to find another path for transmission. If two-hop acknowledgement is not received for any data packet, it is assumed that the data packet is stuck with the void node and then, a new path will be determined and the same data packet is sent with more hops. This process increases energy consumption, but also provides guarantee for successful data delivery. Let the initial energy of each node Eo which is equal to Emin +rand (Erand), where the value of Emin is 60j and that of Erand is 20j. This simulation uses the simplified energy consumption model to transmit the m bits at a distance of k meters. Energy consumption is presented as:

$$E\ tx\ (m\ ,k) = E\ elec * m + E\ am\ p * m * K2 \tag{15}$$

Here, Eelec is the energy required to transmit 1-bit data and Eamp is the acoustic wave attenuation coefficient or acoustic amplifier energy. To receive the m-bit data, energy consumption is presented as:

$$E\ r\ x\ (m) = E\ elec * m \tag{16}$$

The packet delivery ratio also plays an important role in data transmission. Suppose that k1 represents the data packets which are successfully received by the sink node 1, m is the number of sink nodes and n is the number of generated data packets, then packet delivery ratio is defined as:

$$P\ D\ R = (U^{m}{}_{i=1}\ Si)/n \tag{17}$$

According to the energy model, Table 3 shows the simulation parameters.

Table 3. Simulation parameters.

Parameter Name	Parameter Value
Water Surface	1000 * 1000 m
Depth of Water	2000 m
Ex	50W
Erx	158 mW
Number of Sensor Nodes	60-180
Acoustic Propagation Delay	1500 m
Ideal Energy	58 mW
Header Size	88 bits
Payload Size	576 bits
Neighbour Request	48 bits
Acknowledgement	48 bits
Data Rate	16*10^3 bits
Packet Generation Rate	0.2 packet/sec
Weighting Factor α	0.5 range (0,1)

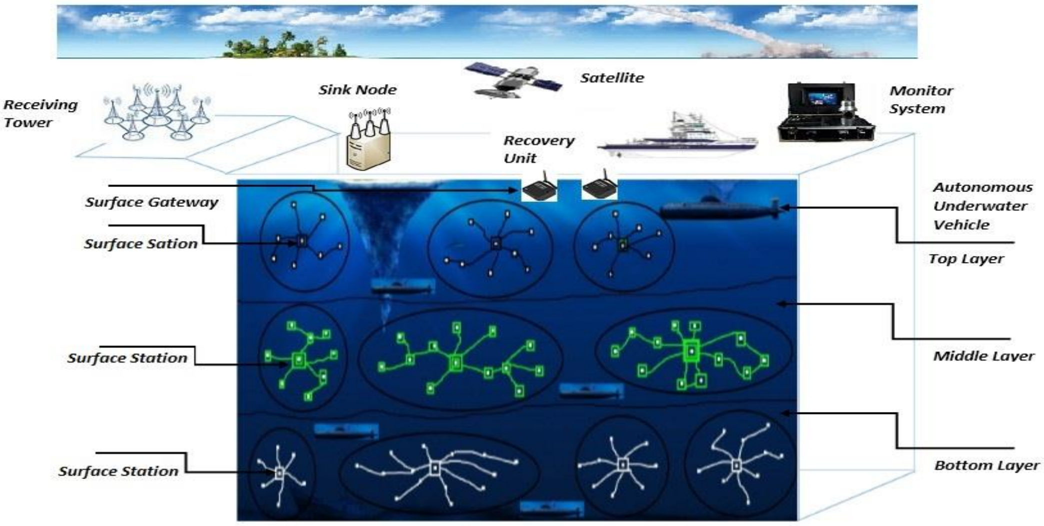


Figure 4. Proposed model for an underwater sensor network.

5. Results and Analysis

The performance of the proposed algorithm is compared with that of the existing approach E2RV. The simulation is done on MATLAB. During simulation, this work considered a three-dimensional area of one kilometer in length and width and two kilometers in depth; i.e., Xmax=Ymax= 1 km and Zmax=2 km. In the simulation, a varying network is considered having 60, 90 and 180 nodes. It is also assumed that sensor nodes are homogenous in terms of transmission range that can vary from 600m to 1000m. Here, source nodes which are also known as data gathering nodes are placed at the bottom of water, while the sink node is placed at water surface. Remaining nodes are randomly deployed at different locations using depth distribution method between source nodes and sink node in the network. These nodes play an important role, as they act as data forwarder node or relay nodes.

In the trial of first simulation, 60 nodes are deployed in the network area randomly except the data source. In the next trial, 120 nodes are deployed, where the locations of the previous 60 nodes are un-altered and only the next 60 nodes are placed randomly in the network. In the proposed scheme, multiple sink nodes are placed at water surface to collect data from others nodes.

5.1 Packet Delivery Ratio

Packet delivery ratio is one of the important performance metrics of any routing or forwarding scheme. Therefore, this work analyzed the PDR performance metric of the proposed technique. The ratio of successfully received data packets by the sink nodes and the total generated packets by the source node is known as packet delivery ratio.

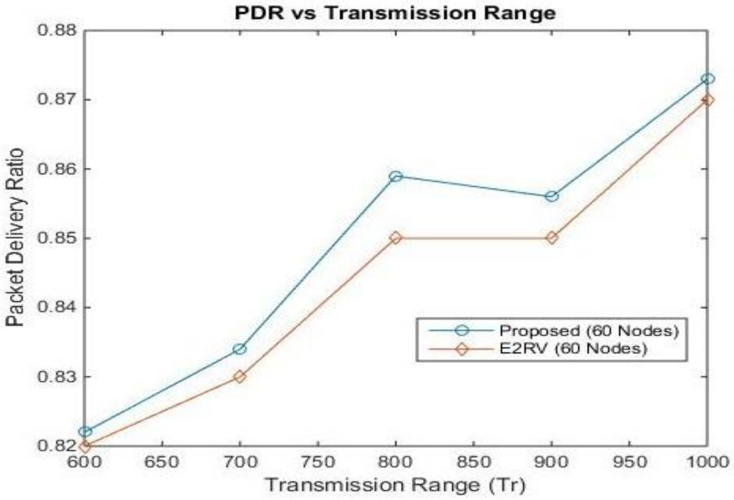


Figure 5. PDR *vs.* transmission range (at 60 nodes).

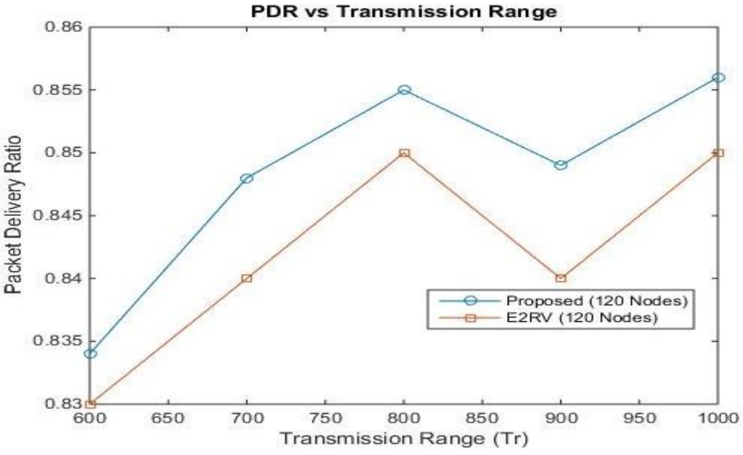


Figure 6. PDR *vs.* transmission range (at 120 nodes).

In case of a fixed network size (like 60 nodes, 120 nodes and 180 nodes) with varying transmission range, Figures 5, 6 and 7 show the improved performance as compared to the E2RV algorithm. As the total number of generated packets is directly proportional to network size, if the data packet generation is high, collision chances increase.

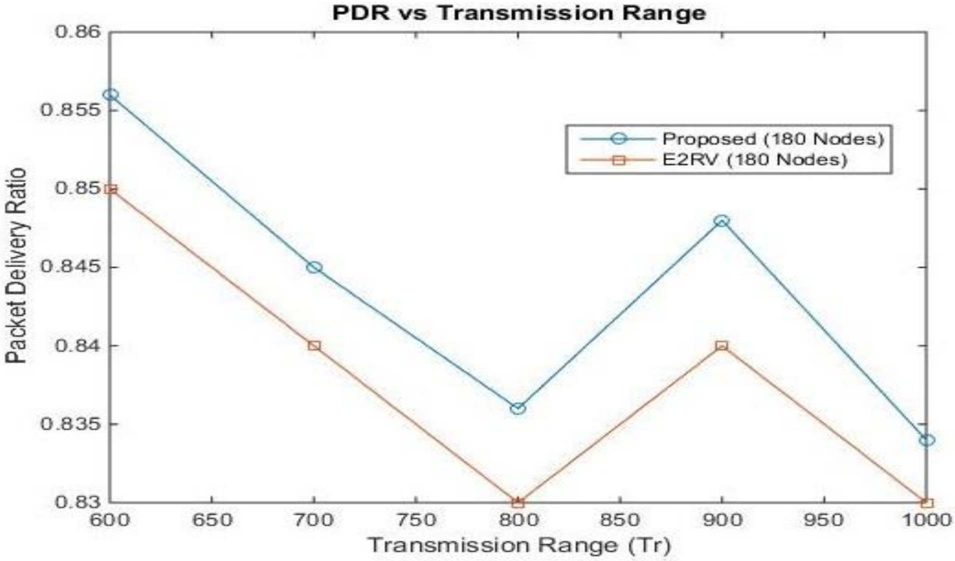


Figure 7. PDR *vs.* transmission range (at 180 nodes).

5.2 Number of Dead Nodes

A sensor node is said to be dead when the battery of the same node is completely used and no power is available to perform the task. The total number of dead nodes depends on the network size. As the network size increases, the number of dead nodes also increases due to the involvement of a large number of nodes to forward the packets. The simulation has been done for varying transmission range with fixed network size in terms of nodes, like 60 nodes, 120 nodes and 180 nodes, respectively. Figures 8, 9 and 10 show that as the number of nodes increases with network size, the number of dead nodes also increases. The reason is that data packet traffic significantly increases with network size.

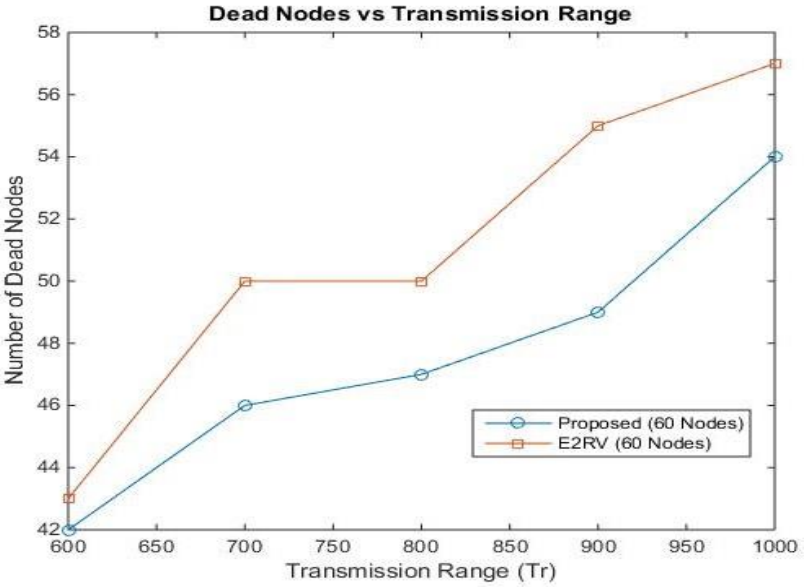


Figure 8. Dead nodes *vs.* transmission range (at 60 nodes).

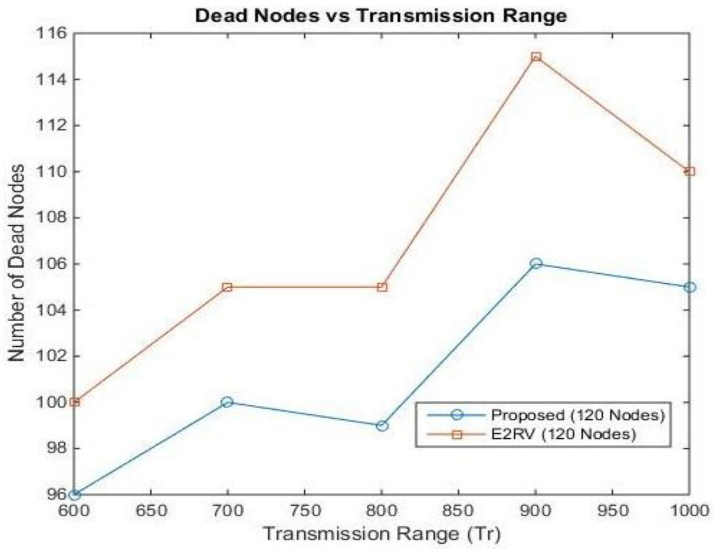


Figure 9. Dead nodes *vs.* transmission range (at 120 nodes).

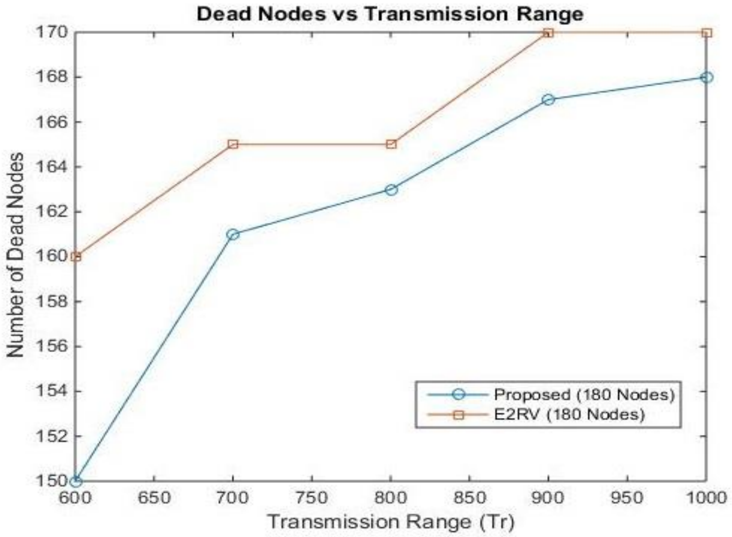


Figure 10. Dead nodes *vs.* transmission range (at 180 nodes).

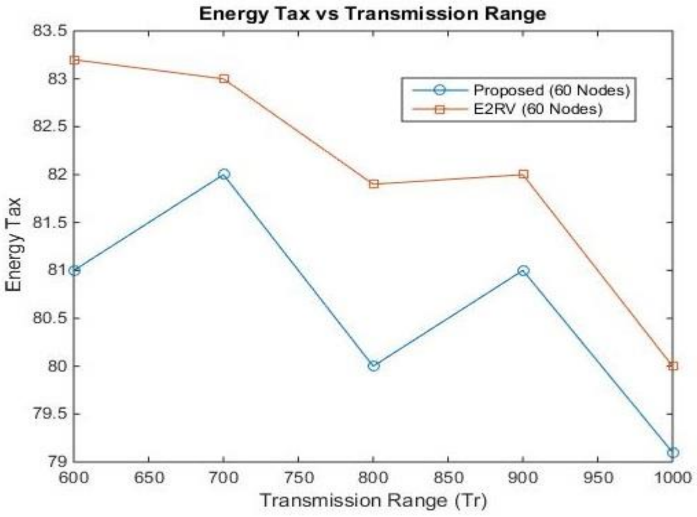


Figure 11. Energy tax *vs.* transmission range (at 60 nodes).

5.3 Energy Tax

The simulation to calculate energy tax has been done with varying transmission range and network size in terms of number of nodes. E2RV consumes more energy and has less packet delivery ratio. When the transmission range (Tr) is varying and network size is fixed, as shown in Figures 11, 12 and 13, respectively, the proposed approach shows good results in terms of energy tax. But, it is also seen that as the network size increases, energy tax slightly increases in both approaches. The reason behind this is that a large volume of energy is consumed to disseminate the large copies of data.

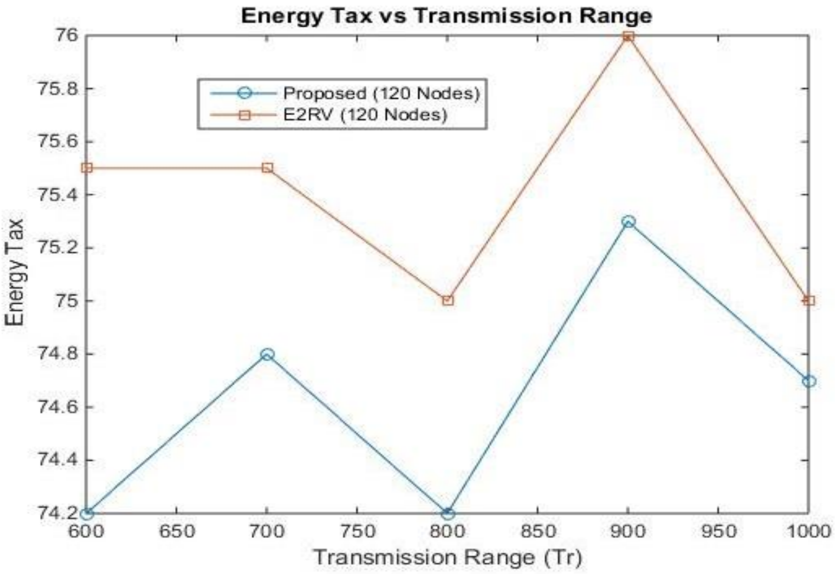


Figure 12. Energy tax *vs.* transmission range (at 120 nodes).

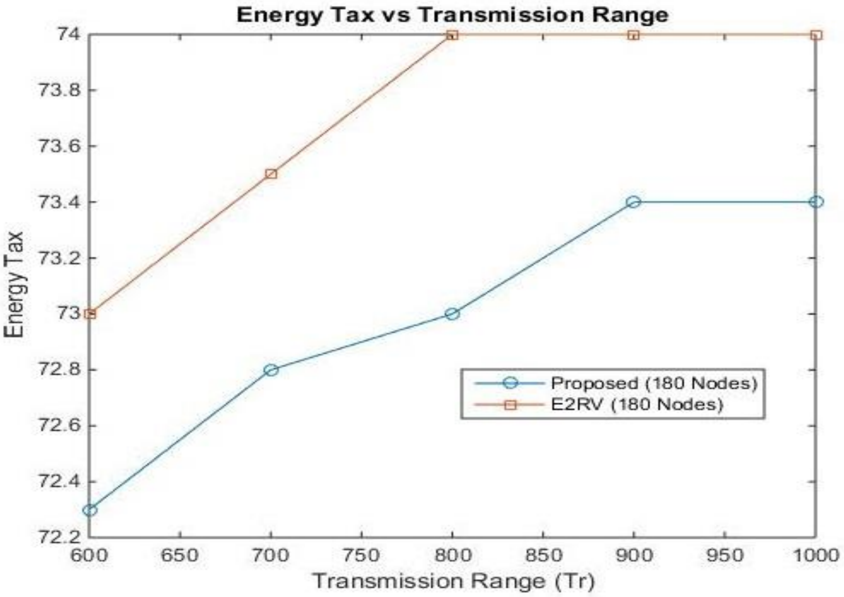


Figure 13. Energy tax *vs.* transmission range (at 180 nodes).

6. CONCLUSIONS

An energy-efficient Quality of Service (QoS) parameter-based void avoidance routing technique for underwater sensor networks has been proposed. In the proposed technique, initially nodes are randomly placed at water surface to form a two-dimensional network topology and a depth optimization technique is proposed here to compute the depth of each sensor node and then inform all the nodes of their optimized depths. By doing this, each node gets its depth and makes a three-dimensional network structure. A candidate forwarder selection method is also implemented here, which is based on fitness function, holding time, depth variance and distance …etc. In addition to this, the proposed work uses the concept of two-hop acknowledgement for successful delivery of data packets at the sink node. The proposed technique has been implemented in MATLAB, where comparison has been done with E2RV approach in terms of Packet Delivery Ratio (PDR), number of dead nodes and energy tax. The proposed approach improves Packet Delivery Ratio (PDR) and reduces the total energy consumption, which in turn decreases the number of dead nodes.

REFERENCES

[1] J.-Y. Lee, N.-Y. Yun, S. Muminov, S.-Y. Shin, Y.-S. Ryuh and S.-H. Park, "A Focus on Practical Assessment of MAC Protocols for Underwater Acoustic Communication with Regard to Network Architecture," IETE Technical Review, vol. 30, no. 5, pp. 375-381, DOI: 10.4103/0256-4602.123119.

[2] A. Khasawneh, M. S. B. A. Latiff, O. Kaiwartya and H. Chizari, "Next Forwarding Node Selection in Underwater Wireless Sensor Networks (UWSNs): Techniques and Challenges," Information, vol. 8, no. 3, 2017.

[3] K. K. Gola and B. Gupta, "Underwater Sensor Networks Routings (UWSN-R): A Comprehensive Survey," Sensor Letters, vol. 15, no. 11, 2017.

[4] R. Zandi, M. Kamarei, H. Amiri and F. Yaghoubi, "Underwater Sensor Network Positioning Using an AUV Moving on a Random Waypoint Path," IETE Journal of Research, vol. 61, no. 6, pp. 693-698, DOI: 10.1080/03772063.2015.1034196, 2015.

[5] A. Muhammad, B. Imran, A. Azween and F. Ibrahima, "A Survey on Routing Techniques in Underwater Wireless Sensor Networks," Journal of Network and Computer Applications, Elsevier, vol. 34, no. 6, pp. 1908-1927, 2011.

[6] M. R. Jafri, S. Ahmed, N. Javaid, Z. Ahmad and R. J. Qureshi, "AMCTD: Adaptive Mobility of Courier Nodes in Threshold-optimized DBR Algorithm for Underwater Wireless Sensor Networks," Proceedings of the IEEE 8th International Conference on Broadband, Wireless Computing, Communication and Applications, IEEE (BWCCA ’13), pp. 93–99, France, 28-30 Oct. 2013.

[7] M. T. Kheirabadi and M. M. Mohamad, "Greedy Routing in Underwater Acoustic Sensor Networks: A Survey," Journal of Distributed Sensor Networks, Vol. 2013, Article ID 701834.

[8] K. K. Gola and B. Gupta, "Underwater Sensor Networks: An Efficient Node Deployment Technique for Enhancing Coverage and Connectivity: END-ECC," International Journal of Computer Network and Information Security (IJCNIS), vol. 10, no. 12, pp. 47-54, 2018.

[9] F. Senel, "Coverage-aware Connectivity-constrained Unattended Sensor Deployment in Underwater Acoustic Sensor Networks," Wireless Communication and Mobile Computing Journal, vol. 16, no. 14, pp. 2052-2064, 2016.

[10] A. Khasawneh, M. S. A. Latiff, H. Chizari, M. Tariq and A. Bamatraf, "Pressure-based Routing Protocol for Underwater Wireless Sensor Network: A Survey," KSII Transactions on Internet and Information Systems , vol. 9, no. 2, pp. 504–527, 2015.

[11] S. Biswas and R. Morris, "ExOR: Opportunistic Multi-hop Routing for Wireless Networks," ACM SIGCOMM Comput. Commun. Rev., vol. 35, pp.133–144, 2005.

[12] T. Javidi and E. Van Buhler, Opportunistic Routing in Wireless Networks, Found. Trends Netw. 2016.

[13] S. M. Ghoreyshi, A. Shahrabi and T. Boutaleb, "An Inherently Void Avoidance Routing Protocol for Underwater Sensor Networks," Proceedings of the IEEE International Symposium on Wireless Communication Systems (ISWCS), pp. 361–365, Brussels, Belgium, 25–28 August 2015.

[14] N. Chakahouk. "A Survey on Opportunistic Routing in Wireless Communication Networks," IEEE Commun. Surv. Tutor., vol. 17, pp. 2214-2241, 2015.

[15] H. Yan, Z. J. Shi and J.-H. Cui, "DBR: Depth-based Routing for Underwater Sensor Networks," Proceedings of the International Conference on Research in Networking, pp. 72–86, Singapore, 2008.

[16] Y. Noh, U. Lee, P. Wang, B. S. C. Choi and M. Gerla, "VAPR: Void-aware Pressure Routing for Underwater Sensor Networks," IEEE Trans. Mobile Comput., vol. 12, pp. 895–908, 2013.

[17] G. A. Hollinger, S. Choudhary, P. Qarabaqi et al., "Underwater Data Collection Using Robotic Sensor Networks," IEEE Journal on Selected Areas in Communications, vol. 30, no. 5, pp. 899–911, 2012.

[18] J.-H. Cui, J. Kong, M. Gerla and S. Zhou, "The Challenges of Building Mobile Underwater Wireless Networks for Aquatic Applications," IEEE Network, vol. 20, no. 3, pp. 12–18, 2006.

[19] F. Emad, S. K. Faisal, Q. M. Umair, S. A. Adil and Q. B. Saad, "Underwater Senosr Networks Application: A Comprehensive Survey," International Journal of Distributed Sensor Networks. vol. 11, no. 11, 2015.

[20] A. Yalcuk and S. Postalcioglu, "Evaluation of Pool Water Quality of Trout Farms by Fuzzy Logic: Monitoring of Pool Water Quality for Trout Farms," International Journal of Environmental Science and Technology, vol. 12, no. 5, pp. 1503–1514, 2015.

[21] P. Xie, H. J. Cui and L. Lao, "VBF: Vector-based Forwarding Protocol for Underwater Sensor Networks," Proc. of the International Conference on Research in Networking, (Networking 2006), Networking Technologies, Services and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communication Systems, pp. 1216–1221, Berlin/Heidelberg: Springer, Coimbra, Portugal, 15-19 May 2006.

[22] M. Ayaz and A. Abdullah, "Hop-by-hop Dynamic Addressing-based (H2-DAB) Routing Protocol for Underwater Wireless Sensor Networks," Proceedings of the International Conference on Information and Multimedia Technology, ICIMT, pp. 436-441, Jeju Island, South Korea, 16-18 December 2009.

[23] N. Chirdchoo, W.-S. Soh and K. C. Chua, "Sector-based Routing with Destination Location Prediction for Underwater Mobile Networks," Proceedings of the 7th IEEE International Conference on Advanced Informaion Networking and Application Workshops, Bradford, UK, 26–29 May 2009.

[24] Md. Ashrafuddin, Md. Manowarul Islam and Md. Mamun-or-Rashid, "Energy-efficient Fitness-based Routing Protocol for Underwater Sensor Networks," International Journal of Intelligent Systems and Applications (IJISA), vol. 5, no .6, pp. 61-69, 2013.

[25] G. Khan and R. K. Dwivedi, "Energy-Efficient Routing Algorithm for Void Avoidance in UWSNs Using Residual Energy and Depth Variance (E2RV)," IJCNC, vol. 10, no. 4, pp. 61-78, July 2018.

[26] E. Isufi, H. Dol and G. Leus, "Advanced Flooding-based Routing Protocols for Underwater Sensor Networks," EURASIP Journal on Advances in Signal Processing, vol. 2016, no. 52, pp. 1–12, 2016.

[27] A. R. Hameed, N. Javaid, S. Islam, G. Ahmed, U. Qasim and Z. A. Khan, "BEEC: Balanced Energy Efficient Circular Routing Protocol for Underwater Wireless Sensor Networks," Proceedings of the 8th IEEE International Conference on Intelligent Networking and Collaborative Systems, Ostrava, Czech Republic, 7–9 September 2016.

[28] A. Sher, N. Javaid, G. Ahmed, S. Islam, U. Qasim and Z. A. Khan, "MC: Maximum Coverage Routing Protocol for Underwater Wireless Sensor Networks," Proceedings of the 19th IEEE International Conference on Network-based Information Systems, Ostrava, Czech Republic, 7–9 September 2016.

[29] Z. Rahman, F. Hashim, M. F. A. Rasid and M. Othman, "Totally Opportunistic Routing Algorithm (TORA) for Underwater Wireless Sensor Network," PLoS ONE, vol. 13, no. 6, [Online], Available: https://doi.org/10.1371/journal.pone.0197087, 2018.

[30] S. H. Bouk, S. H. Ahmed, K.-J. Park and Y. Eun, "EDOVE: Energy and Depth Variance-based Opportunistic Void Avoidance Scheme for Underwater Acoustic Sensor Networks," Sensors, vol. 17, no. 10, 2017.

[31] C.-J. Huang, Y.-W. Wang, H.-H. Liao, C.-F. Lin, K.-W. Hu and T.-Y. Chang, "A Power-efficient Routing Protocol for Underwater Wireless Sensor Networks," Applied Soft Computing, vol. 11, no. 2, pp. 2348–2355, 2011.

[32] P. Xie, Z. Zhou, Z. Peng, J.-H. Cui and Z. Shi, "Void Avoidance in Three-dimensional Mobile Underwater Sensor Networks," Proceedings of the International Conference on Wireless Algorithms, Systems and Applications, pp. 305–314, Boston, MA, USA, 16–18 August 2009.

ملخص البحث:

أصبحت شبكات المجسات تحت الماء أحد أكثر مجالات البحث إثارة؛ فقد فتحت الباب واسعاً أمام الباحثين لإجراء دراسات متخصصة في هذا المجال. والجدير بالذكر أن هناك العديد من المشكلات المرتبطة بشبكات المجسات تحت الماء، لعل من أبرزها منطقة الفراغ التي تحطّ من أداء الشبكة. وهي مشكلة تنطوي على عدم وجود أي عقدة ناقلة تنقل حزم البيانات من عقدة الى أخرى.

لذا يتلخص غرض هذا البحث في تجنب منطقة الفراغ. ولتحقيق ذلك الغرض، تستخدم هذه الدراسة تقنية فعالة من حيث استهلاك الطاقة، تقوم على عدد من متغيرات جودة الخدمة؛ من أجل تجنب الفراغ في شبكات المجسات تحت الماء. ولتجنب منطقة الفراغ في تلك الشبكات، تستخدم التقنية المقترحة معلومات العقد ذات القفزتين. أما منغيرات جودة الخدمة التي أُخذت بعين الاعتبار، فهي: الاحتفاظ، والطاقة المتبقية؛ من أجل إيجاد العقدة الناقلة المُثلى التي ستنقل حزم البيانات الى غايتها.

لقد طُبقت الخوارزمية المقترحة على "ماتلاب"؛ إذ بينت النتائج أداءً أفضل للتقنية المقترحة من حيث: معدل تسليم الحُزم، واستهلاك الطاقة، وعدد العُقد الميتة، وذلك بمقارنتها بالتقنية المسمّاة (E2RV) التي يشيع استخدامها في شبكات المجسات تحت الماء.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 05, No. 03, December 2019.

THE UTILIZATION OF EEG SIGNAL IN VIDEO COMPRESSION

Qasem Qananwah[1], Hussein Alzoubi[2], Ruba Banimfarij[2], Ahmad Dagamseh[3] and Oliver Hayden[4]

(Received: 27-Jul.-2019, Revised: 6-Oct.-2019, Accepted: 21-Oct.-2019)

ABSTRACT

Due to technology advances in multimedia, larger storage spaces, large internet bandwidth and high-transmission speed are required for the transmission of videos. Video compression techniques play a vital role in reducing video size; therefore, smaller storage space and lower internet bandwidth are eventually required. In this paper, the EEG signal is used to modify the compression ratio of videos based on the interest of the viewer. This is performed by associating the compression ratio applied to the video with the degree of interest using a group of frames. This interest for a group of frames is measured using the EEG signal to demonstrate the viewer responses to videos. Statistical techniques applied to the EEG signal (such as peaks-over-threshold and time-of-peaks-over-thresholds) are used to extract the frames of interest. Peak signal-to-noise ratio (PSNR), Structural Similarity Index (SSIM) and Mean-Square Error (MSE) are used to compare the performance of the proposed technique with the MPEG-4 technique. The results show a reduction of 15 % in the video size compared with the MPEG-4 technique without deteriorating the quality of the videos.

KEYWORDS

Video compression, EEG signal, MPEG-4, PSNR, SSIM, Brainwaves.

1. INTRODUCTION

Nowadays, multimedia plays an essential role in various human activities, such as learning, leisure and communication. High-definition videos require large bandwidth and storage space. Therefore, video compression techniques are used to minimize the number of bits to represent data. This results in more free storage capacity, rapid file transfer and efficient use of bandwidth when compared with uncompressed versions.

Video compression can be categorized into lossy or lossless compression. Lossy compression results in high compression ratio (when unnecessary information is removed) and lower bits with an acceptable level of quality. However, the original data cannot be accurately recovered [1]. Lossless compression results in complete recovery of the original data with lower compression ratio. International Telecommunications Union (ITU) and the International Standards Organization/International Electrotechnical Commission (ISO/IEC) developed the standards of video compression. Examples of these international standards are MPEG [2] and H.263 [3]. MPEG or MPEG-1 is the first true multimedia standard that has specifications for coding compression, transmission of audio, transmission of video and data streams in a series of synchronized mixed packets. MPEG-2 is used to perform high-quality transmission, multi-channel and multimedia over a broadband network like ATM. MPEG-4 provides high-compression characteristics of MPEG [4]. MPEG-4 supports all features of MPEG-1 and MPEG-2 and supports lower bandwidth-consuming applications (e.g. mobile phones). It is mainly utilized in digital television, interactive graphics applications (synthetic content) and the World Wide Web [5]. Recent techniques have emerged to target improving compression of high-dimensional or new video formats based on perceptual coding. Ki et al. presented a novel discrete cosine transform (DCT) model and applied it for perceptual video coding (PVC) [6]. The model was applied to high-efficiency video coding (HEVC). To further enhance the compression performance based on perceptual video coding, Prangnell and Sanchez proposed a novel JND-based PVC method to reduce the bit rate [7]. For

1. Q. Qananwah is with the Department of Biomedical Engineering, Yarmouk University, Irbid, Jordan. Email: Qananwah@gmail.com
2. H. Alzoubi and R. Banimfarij are with the Department of Computer Engineering, Yarmouk University, Irbid, Jordan. Emails: halzoubi@yu.edu.jo and Banimfari@yahoo.com
3. A. Dagamseh is with the Department of Electronic Engineering, Yarmouk University, Irbid, Jordan. Email: ahmad.dagamseh@gmail.com
4. O. Hayden is with the Department of Electrical and Computer Engineering, Technical University of Munich, Munich, Germany. Email: oliver.hayden@tum.de

optimizing video streaming, Takeuchi et al. introduced a perceptual video quality encoding solution [8]. They developed an estimator utilizing support video regression (SVR) to fulfill this function. Helmrich et al. proposed a simple algorithm, based on the human visual system, using perceptual video coding QPA [9]. As a result of the subjective tests run, the QPA was adopted in VTM (VVC). There are also other techniques aiming to improve the compression performance based on control of bit rate. Perez-Daniel and Sanchez proposed a multi-R - λ-model approach [10]. Using the peak signal-to-noise ratio (PSNR) metric, the results show that their approach is comparable with current RC techniques used in HEVC.

The High-Efficiency Video Coding (HEVC), aka H.265, is the latest video coding standard, which is a joint project of the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG) [11]. H.265 is the successor of H.264 (MPEG-4, Part 10). We would like to point out that HEVC is also applicable for our work, since the core spatial and temporal compression techniques of H.265 are the same as those used in its predecessors. The newest coding techniques, such as scalable video coding and 3-D/stereo/Multiview video coding [11] that were added to H.265, are irrelevant to our work in this paper.

Electroencephalography (EEG) is a technique used to record the electrical activity of the brain at different positions on the scalp [12]. The recorded activities represent the voltage potential induced due to the ionic motion within the neurons in human brain cells [13]. EEG signal plays a key role in many applications, e.g. diagnosing patients for different disorders, like a seizure, neuromarketing, brain-computer interfaces (BCIs), clinical and psychiatric studies [14].

One of the modern applications of the EEG signal is image and video compression [15]-[16]. The information in the EEG signal can be utilized to improve the compression ratio, especially in the era of the internet. Few researchers have addressed this topic; however, the published research has mainly focused on still images. Li et al. found that 3D visual fatigue can be detected using the EEG signal [17]. Lea Lindemann and Marcus Magnor concluded that the EEG signal might be used as a tool for evaluating image quality, where much information is implicitly encrypted in the signal [16]. Evaluating images using the features of the EEG signal improves image quality [18]. The utilization of the features of EEG signals may also be considered as an ultimate goal to produce more efficient video-compression algorithms. Tan et al. proposed a deep transfer learning algorithm suitable for knowledge transfer. Their approach was approved by applying their algorithm in EEG classification tasks [19].

The use of EEG or BCI in multimedia has made some progress in recent years. Engelke et al. presented a review of psychophysiology-based assessment for Quality of Experience (QoE) in multimedia [20]. Furthermore, Bosse et al. presented and discussed different approaches for multimedia quality assessments using BCI and addressed the challenges relevant to its community [21]. Additionally, Bosse et al. [22] evaluated still images using steady-state evoked potential (SSVEP). They used SSVEP, which represents neural response (the EEG signal), to assess the quality of texture images. Avarvand et al. presented a study for quality assessment of stereoscopic images using the EEG signal [23]. They measured the event-related potential (ERP) for 2D and 3D images and analyzed their measurements using time domain and frequency domain. They reported an increase in the amplitude of 3D images compared to 2D images. Many other studies in the literature have addressed image-quality assessment using various techniques, e.g. ERP [24]-[26]. In addition, Bosse et al. presented a video-quality assessment based on psycho-physiological techniques [27]. The EEG signal has been also considered in the literature to measure and determine video quality [28]-[30]. Although techniques related to assisting video quality using the EEG signal are not well developed, it is promising that these techniques may attain comparable results in the near future.

Video compression is a very helpful technique to facilitate video transmission, as the size is reduced, thereby preserving the bandwidth and increasing the transmission speed. Currently, the algorithms used in video compression (ex. MPEG) utilize a fixed compression ratio for the entire video. Generally, modifying the compression ratio of the video based on the Region-of-Interest (ROI) frames will improve the compression efficiency. This can be performed by determining the interesting frames in the videos and manipulating the compression ratio accordingly (i.e., increasing the compression ratio for the frames of less interest and *vice versa*). In this paper, the EEG signal is used to extract the interests of the video viewer by means of feature extraction and utilize them to manipulate the rate of video compression. The video is projected in the EEG signal as an evoked potential, which varies according to the interests of the viewer. The correlation between the features of the EEG signal (as the response of the video viewer) and the amount of compression will be accordingly identified.

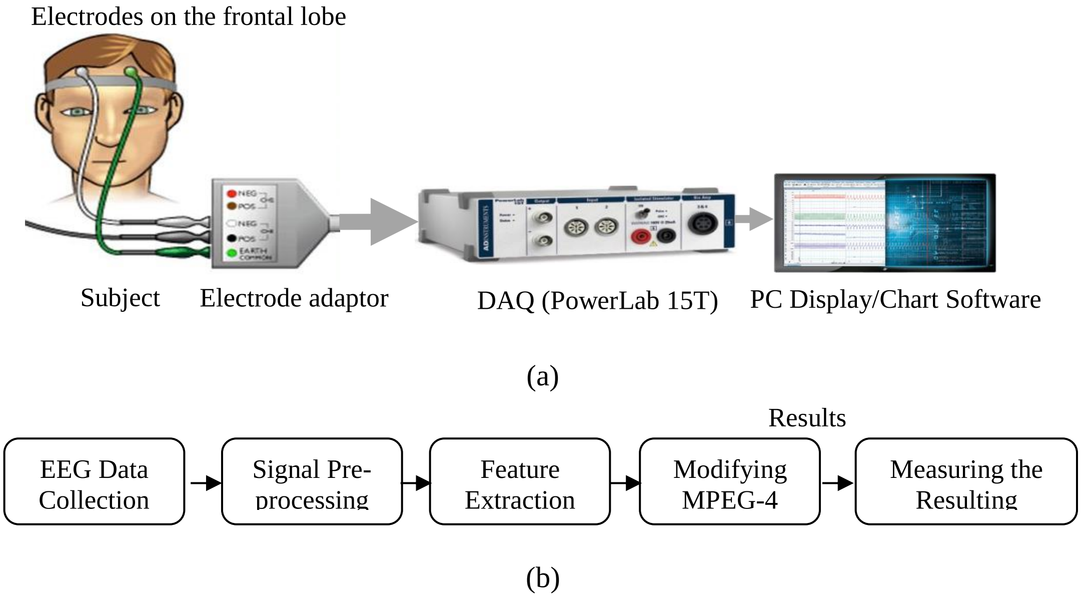
2. Materials and Methods

2.1 Measurement Setup

Measurements were performed using a hardware system equipped with software provided by AD Instrument system (i.e., PowerLab 15T). The EEG signal was measured using electrodes placed at the frontal lobe, as shown in Figure 1(a). The electrode-skin impedance was reduced using gel (paste) to the range of 5 kΩ for the best EEG signal quality. Various tests were used to ensure the proper signal-to-noise ratio, as the EEG signal is a random signal. These tests were performed at the beginning of each measurement. The tests performed with the EEG signal are:

- Eye blinking test: This test includes determining the existence and number of the eye blinking artifact in the EEG signal when the subject blinks repeatedly;
- Alpha-wave test: This test includes requesting the subject to close his/her eyes and be calm without any brain activity. The EEG signal variations (existence of the alpha-wave) can be noticed when the subject opened his/her eyes;
- Clenching teeth test: This test includes clenching the teeth and accordingly, the amplitude and the frequency of the EEG signal increase.

The EEG signals were measured while the subjects were watching the videos. The signals were stored through the input terminals at the front panel of the setup. The EEG signals were recorded using two electrodes with about 2 cm above the hairline and 5 cm separation. A third electrode was placed on the earlobe or clavicle (i.e., ground reference electrode). The onset of the measurements was synchronized with the start of the video. A data acquisition system (DAQ) was used to transfer the measured signals to the computer with a sample rate equal to 1 kSPS. Afterwards, the signals were handled according to the procedure illustrated in Figure 1(b).



(a)

(b)

Figure 1. (a) AD Instrument acquisition system used in the measurement and (b) a block diagram of the signal processing procedure applied to the EEG signals.

2.2 Video Characteristics

Three uncompressed muted video-clips were selected carefully, which include some important stimuli, such as rapid motion, luminance and color variation. The test videos were uncompressed videos and are usually used in compression evaluation in typical studies. In spite of the short duration of these videos, their characteristics contain the main features that are needed for the evaluation of the proposed compression techniques. The videos were muted to measure the participant's attention according to the video features only.

The first video contains a colorful Chinese city with many details and vehicles crossing the camera in various directions. Such a video test guarantees the existence of two important parameters (i.e., motion vector and color variation). It involves local motion (motion of objects; i.e., cars), while the global motion (camera motion) was relatively small. On the other hand, the spatial information was large, as the frames were colorful.

The second video was about sea waves hitting the coast with a slow movement in the camera. This involves the presence of the bluish color with the small motion vector. It involves less local motion (motion of objects; i.e., sea waves), small global motion (camera motion) and very limited spatial information, as the frames contain small color variations (i.e., mainly blue).

The third video was about two guys playing table tennis. The motion is fast and thus the motion vector is large due to the sudden change in the players' position. It involves local motion (motion of objects; i.e., athletes), while the global motion (camera motion) was relatively small. On the other hand, the spatial information was relatively small.

The EEG features have been extracted from each subject for each video to evaluate the degrees of the subjects' attention and interest to each video segment (i.e., a group of frames). Video details are shown in Table 1, while Figure 2 shows a sample image of one of the videos used in this study.

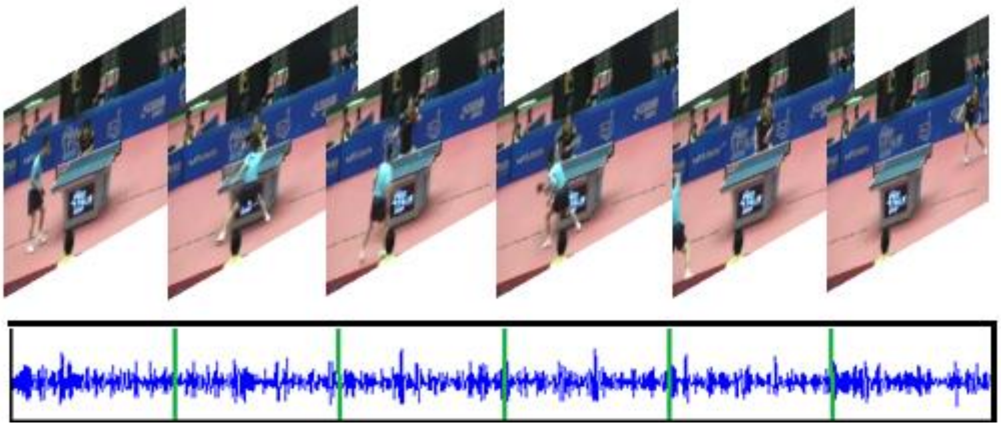


Figure 2. Sample images of one of the test videos used in this study.

Table 1. Details of the test videos.

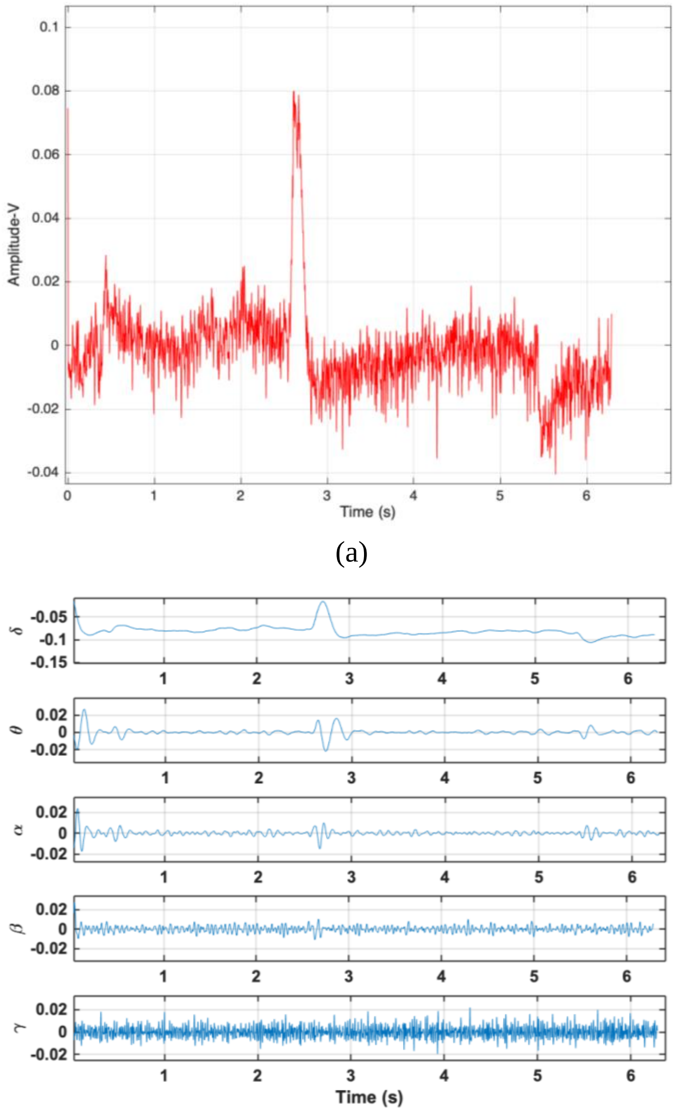
Video	Size (kB)	Resolution	Number of Frames
1	13609	1280×720	460
2	13185	1280×720	300
3	13100	320×240	1200

2.3 EEG Data Collection

EEG data signals were collected from twenty subjects who participated in the experiments. The subjects were distributed as follows; eight males, nine females (with ages between 18 and 30 years) and three children (with ages between 8 and 12 years). During the measurements, the subjects were asked to be calm with reduced movement and minimum possible eye blinking. They were seated in a comfortable position on a chair with none of the electrodes placed on the subjects' heads. Once the subjects got familiar with the surroundings, the electrodes were connected to their positions. Figure 3 (a) shows a sample of the raw EEG signal acquired in this study with its decomposed components in Figure 3 (b).

2.4 Signal Processing

MATLAB environment was used for performing signal processing on the measured EEG signals. Signal processing involved filtration, frequency conversion and feature extraction. Power-line noise appeared



(a)

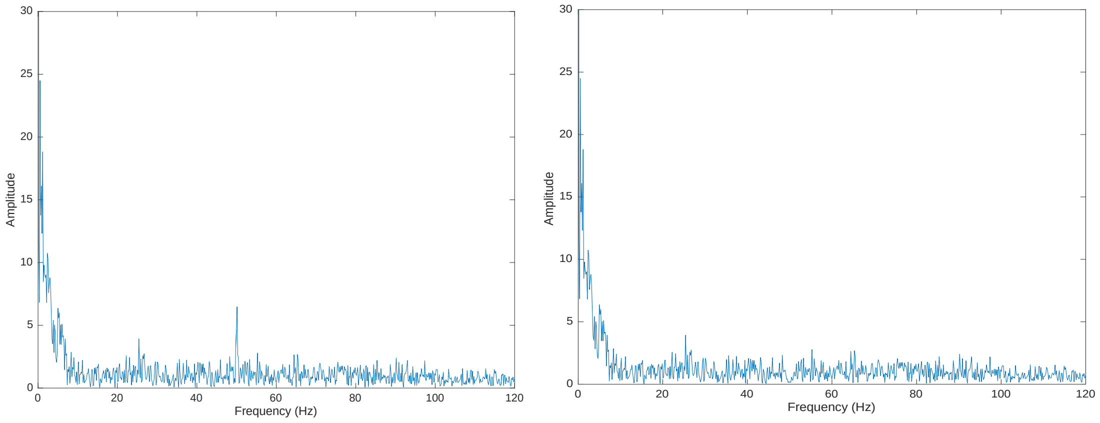
(b)

Figure 3. A sample of the EEG signal: (a) the raw signal and (b) the decomposed and processed components.

in the measured signals, which was removed with a notch-filter with a narrow frequency-band (49-51 Hz). Figure (4) shows the signal before and after the filtration process. High-frequency noise was also removed using a low-pass filter with a corner frequency of 100 Hz. The other sources of noise which normally appear in any EEG signals, like human artifact (low frequency) and electromagnetic noise superimposed on EEG, EMG and EOG signals, did not appear due to the careful procedure followed to reduce sources of noise. The EEG signals were decomposed into their corresponding waves. Beta-wave in the EEG signal has been used in the analysis, since it is registered during high activity. Beta-wave exists when the subject is concentrating and while performing brain activities (i.e., thinking or during visual stimuli) [31]-[33]. The extracted features show that there is a brain activity represented by the following features (i) Average of peaks-over-threshold (ii) Average duration of peaks over threshold and (iii) Power spectrum of the signals. Those features were chosen, as they provide a good indication of the brain activity, which is directly related to the stimulus. The peaks-over-threshold values were selected according to the central-limit theorem. Each segment is composed of 8 frames.

2.5 Feature Extraction

The EEG signal was used to extract the interesting frames in the video for the viewers using statistical analysis techniques. The statistical distribution of the EEG signal was calculated and a threshold was obtained above 5 % of the distribution. The amplitudes above the threshold with a certainty of 95% were



(a) (b)

Figure 4. An example of a frequency-domain representation of the EEG signal: (a) raw signal and (b) filtered signal.

calculated to denote a higher activity in the brain due to stimulus (i.e., video frames). The time of these peaks above the threshold was also determined. The average peaks and average times were used to locate the region of interest (ROI).

As an essential requirement for feature extraction, the onset of the video was synced with the recording of the EEG signal. The EEG signals were decomposed into their corresponding brain-waves. Data analysis was performed on the Beta-wave (i.e., frequency contents of 8 Hz to 14 Hz), which appears during high-mental activities. Therefore, in the following text, wherever the EEG signal is used, the Beta-wave is meant to be. For each frame, the length of the EEG signal was determined (according to Equation (1)) and afterwards, the segment of the signal for each group of frames was determined. Within the MPEG-4 coded video stream, the Group of Pictures (GOP) was chosen to be 8 with a 34 Quantization Parameter (QP) in the frames' analysis. The applied steps for the proposed technique were as follows:

1. The length of the EEG signal for each frame was calculated using Equation (1).

$$L_{EEG}=N/T \tag{1}$$

where L_{EEG} is the length of the EEG signal per frame (i.e., the EEG signal segment), N is the length of the entire EEG signal and T is the number of video frames.

2. Each frame corresponding to an EEG sequence was computed according to Equation (2).

$$EEG_{segment(i)} = v*L_{EEG} \tag{2}$$

where v is the index of the frame.

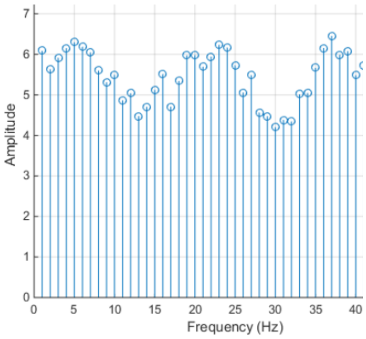


Figure 5. An example of EEG mean value for each segment distributed along the frequency range.

Due to the random property of the EEG signal, each EEG sequence was represented by one value (i.e., the mean value of the amplitude of segments). Figure 5 shows the corresponding averaged EEG signal for each video segment.

2.6 Modification of MPEG-4 Algorithm

The MPEG-4 technique is modified for each frame, where the averaged values of the EEG signal were used to choose the I-frames, P-frames and B-frames of MPEG-4 video. The threshold values were selected according to the EEG signal frequency range and the brain-wave associated with it. The frames with the maximum average amplitudes and maximum average times were correlated with the region of activation (in the EEG signal), in such a way that the first frame in this sequence represents the I-frame. The thresholds used to find the I-frame were determined experimentally, so that the video quality was assured, represented by the PSNR and Structural Similarity Index (SSIM) parameters. The best threshold that matches the previous criterion was found to be 0.8 for the I-frame and 0.6 for the P-frame, while the threshold for the B-frame was <0.6.

3. Results and Discussion

The efficiency of the proposed approach was evaluated and compared with the MPEG-4 standard approach through (i) comparing the video size before and after the compression process (ii) calculating the PSNR and (iii) calculating the SSIM and comparing it with that of the original video. PSNR, SSIM and Mean-Square Error (MSE) parameters were used in the evaluation of the compression technique as an indication of image quality and to predict the human visual response. Therefore, PSNR and SSIM were used for making the comparison between the proposed method and the MPEG-4. The results show that the proposed method does not deteriorate the quality of images. PSNR (in dB) was computed according to Equation (3).

$$PSNR = 10 \times log_{10}\left(\frac{MAXI^2}{MSE}\right) \tag{3}$$

where, *MAXI* is the maximum pixel value of each frame and MSE is:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 \tag{4}$$

where, I is the original uncompressed frame and K is the coded video frame; i and j are the row and column pixel index, respectively. SSIM can be calculated using Equations (5-9) with SSIM equals one if the two images are identical [34]:

$$\mu_x = \frac{1}{T}\sum_{i=1}^{T} x_i \tag{5}$$

$$\mu_y = \frac{1}{T}\sum_{i=1}^{T} y_i \tag{6}$$

$$\sigma_x^2 = \frac{1}{T-1}\sum_{i=1}^{T}(x_i - x)^2,\ \sigma_y^2 = \frac{1}{T-1}\sum_{i=1}^{T}(y_i - y)^2 \tag{7}$$

$$\sigma_{xy}^2 = \frac{1}{T-1}\sum_{i=1}^{T}(x_i - x)^2(y_i - y)^2 \tag{8}$$

$$(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{9}$$

where, c_1 and c_2 are stabilizing coefficients.

The MPEG-4 algorithm was modified by varying the compression ratio based on the degree of interest instead of using a constant compression ratio. The portion of the videos with higher interest was less compressed compared with the portions with less interest. The videos were shown two times: trial version and testing version for each subject. Repeating the measurements during the experiments didn’t influence the results in terms of the features extracted from both trials. For the first video, the size of the uncompressed version was 13609 KB, while for the conventional MPEG-4 compression, it was 7173 KB (see Table 1). Applying the modified MPEG-4 algorithm on the uncompressed version of the first video revealed that the average video size has reduced to 5992.2 KB. Table 2 summarizes the size of the compressed video using the proposed approach (i.e., modified MPEG-4 algorithm) for each participant (P). The results show clearly that the size of the compressed video is lower compared with the conventional MPEG-4 algorithm. The quality of the video has not deteriorated significantly by using the modified MPEG-4 algorithm, as indicated by the PSNR, SSIM and MSE results. These results are obtained by the measurements performed on all participants.

Table 2. Compression of: (a) first (b) second and (c) third test video using the proposed method for different participants.

(a)		**(b)**		**(c)**	
Sample	**Video size (KB)**	**Sample**	**Video size (KB)**	**Sample**	**Video size (KB)**
P1	6021	P1	5829	P1	6021
P2	5973	P2	5733	P2	5973
P3	6117	P3	5832	P3	6069
P4	6021	P4	5829	P4	6021
P5	5925	P5	5877	P5	6069
P6	5973	P6	5781	P6	5973
P7	5925	P7	5733	P7	6069
P8	6021	P8	5877	P8	6117
P9	6165	P9	5781	P9	5925
P10	5877	P10	5829	P10	5973
P11	6069	P11	5877	P11	6165
P12	6021	P12	5781	P12	6213
P13	5973	P13	5733	P13	6261
P14	5877	P14	5829	P14	6117
P15	6021	P15	5684	P15	5877
P16	6069	P16	5532	P16	6021
P17	5877	P17	5877	P17	5973
P18	5925	P18	5829	P18	6117
P19	6021	P19	5781	P19	6060
P20	5973	P20	5829	P20	6261

The same procedure has been performed for the second and third test videos. The sizes of the uncompressed versions were 13185 KB and 13100 KB, while for the conventional MPEG-4 compression, these sizes were 7074 KB and 6933 KB for the second and third test videos, respectively (see Table 2). Applying the modified MPEG-4 algorithm on the uncompressed version of the videos resulted in reducing the average video sizes to 5792.65 KB and 6063.75 KB for the second and third videos, respectively.

Based on these results, the average difference between frames for the three videos was calculated and found to be 2.1242 KB, 1.1258 KB and 2.4242 KB, respectively. The large difference in video frames indicates low compression ratio and *vice versa*; the low difference in video frames represents high compression ratio. All the three test videos have shown a possibility to increase the compression ratio without deteriorating significantly the quality of the video (i.e., decrease the videos sizes with 1180.8, 1281.35 and 869.25 KB, respectively, compared with MPEG-4 method).

PSNR is usually used to measure the quality of reconstruction of lossy compression codecs. To assess the quality of the compressed videos using the proposed approach, PSNR was determined and compared with PSNR for the same videos using the MPEG-4 method. All the results of the three test videos using the modified MPEG-4 compression approach show low variations in the average value of PSNR when compared with the typical MPEG-4 method. Figures 6 and 7, respectively, show examples of the frame-by-frame PSNR results for the compressed videos using MPEG-4 and the proposed technique for the first video (see Table 3).

SSIM is an image quality assessment method used to measure the similarity between two images. For the compressed video and the original video, SSIM was determined for the MPEG-4 method and the proposed approach. The results show that the compressed videos using the proposed approach have comparable similarity values compared with the MPEG-4 approach (see Table 4). The difference between the two methods was less than 2% in the worst scenario for all participants.

Table 3. Average PSNR of the tested videos using the MPEG-4 method compared with the proposed approach.

Video	Average PSNR (dB)	
	MPEG-4 Method	The Proposed Method
1	46.90	40.69
2	48.74	45.13
3	46.87	45.50

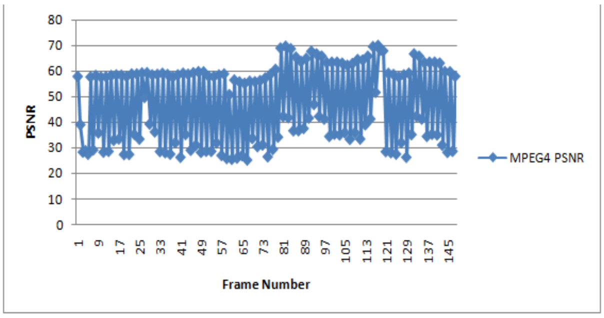


Figure 6. PSNR for the first video using MPEG-4 compression.

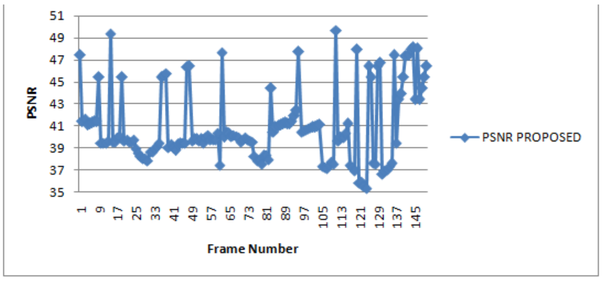


Figure 7. PSNR for the first video using the proposed approach.

Table 4. Average SSIM of the tested videos using the MPEG-4 method compared with the proposed approach.

Video	Average SSIM (%)	
	MPEG-4 Method	The Proposed Method
1	91.81	90.75
2	94.61	92.64
3	93.25	92.33

MSE has been used to check the performance of the proposed compression technique relative to the standard MPEG-4 technique. The results show that there is no significant difference between the video played with the MPEG-4 and the same video compressed with the proposed technique. For the three test videos, compared with the MPEG 4 videos, the averaged absolute differences between frames using the proposed technique were found to be 1.7259, 1.200, 1.9856, respectively. Figures (8-10) show examples of different frames for the MPEG-4 videos compared to the videos with the proposed technique together with the MSE graphs for each frame.

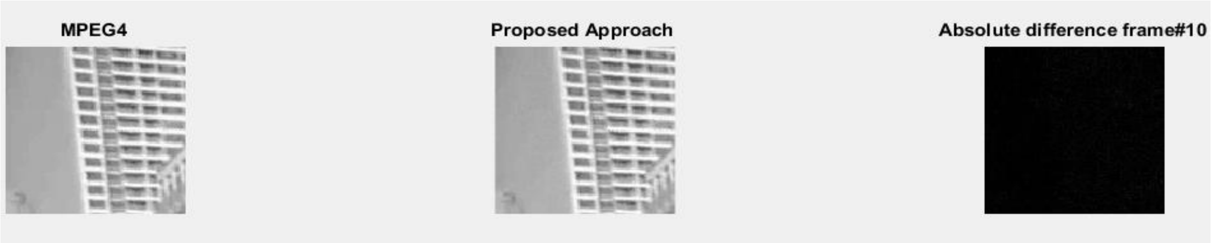


Figure 8. The absolute difference in the first video for frame number 10.

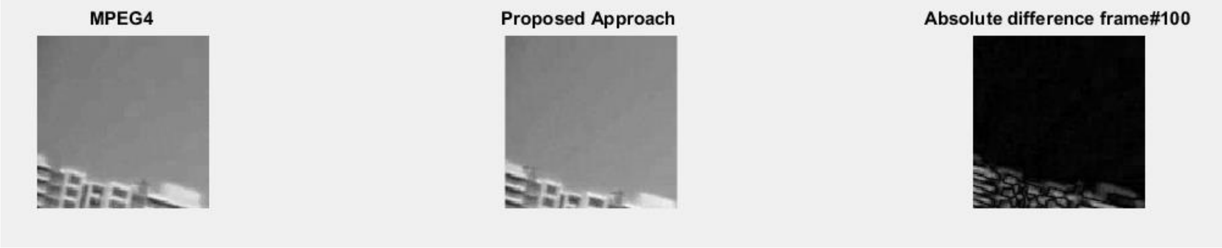


Figure 9. The absolute difference in the first video for frame number 100.

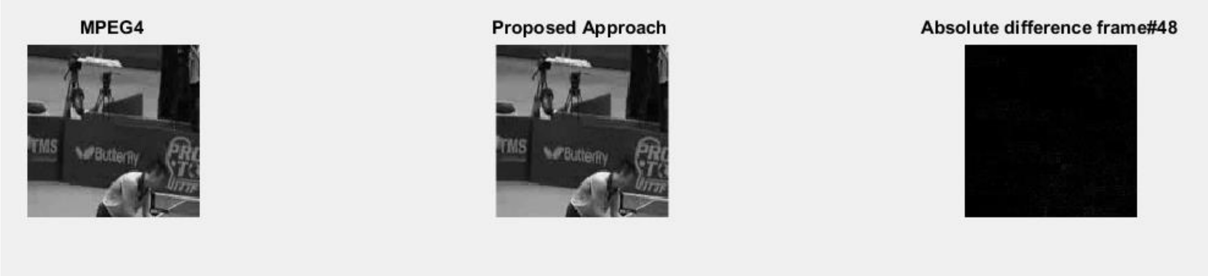


Figure 10. The absolute difference in the third video for frame number 48.

4. CONCLUSIONS

Videos require a large space for storage in addition to a wide internet bandwidth, which necessitates an efficient video compression technique. Electroencephalogram (EEG) signals can be used to measure the responses of video viewers and their interest in different segments of the video. Since the interest through the video varies, an adjustable compression ratio could be applied (i.e., used by conventional compression algorithms). Therefore, the EEG signal can be used to adapt the compression rate. The MPEG-4 was modified to comprise an adjustable compression ratio, so that a high compression ratio is applied to the frames with low interest and a low compression ratio is applied to the frames with high interest. The results obtained illustrate a possibility to reduce the size of the compressed video based on the video content and achieve higher compression ratios. The features of the EEG signal, which represent the viewer’s attention, were used to compress videos to a lower size than in the MPEG-4 technique. In terms of video quality and based on PSNR, SSIM and MSE parameters, the proposed approach showed a comparable quality to the MPEG-4 technique.

REFERENCES

[1] N. Memon and K. Sayood, "Lossless Compression of Video Sequences," IEEE Trans. Commun., vol. 44, no. 10, pp. 1340–1345, 1996.

[2] O. Avaro, A. Eleftheriadis, C. Herpel, G. Rajan and L. Ward, "MPEG-4 Systems: Overview," Signal Process. Image Commun., vol. 15, no. 4–5, pp. 281–298, 2000.

[3] K. Rijkse and K. Research, "H.263: Video Coding for Low-Bit-Rate Communication," IEEE Communications Magazine, pp. 42–45, 1996.

[4] S. Ponlatha and R. S. Sabeenian, "Comparison of Video Compression Standards," Int. J. Comput. Electr. Eng., vol. 5, no. 6, pp. 549–554, 2013.

[5] J. G. Webster, Medical Instrumentation: Application and Design, 4th Ed., Wiley, 2010.

[6] S. Ki, S. H. Bae, M. Kim and H. Ko, "Learning-based Just-noticeable-quantization-distortion Modeling for Perceptual Video Coding," IEEE Trans. on Image Processing, vol. 27, no. 7, pp.3178-3193, 2018.

[7] Prangnell, Lee and V. Sanchez, "JND-based Perceptual Video Coding for 4:4:4 Screen Content Data in HEVC," Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1203-1207, 2017.

[8] M. Takeuchi, S. Saika, Y. Sakamoto, T. Nagashima, Z. Cheng, K. Kanai, J. Katto, K. Wei, J. Zengwei and X. Wei, "Perceptual Quality Driven Adaptive Video Coding Using JND Estimation," Picture Coding Symposium (PCS), 179-183, 2018.

[9] Helmrich, R. Christian et al., "Perceptually Optimized Bit-allocation and Associated Distortion Measure for Block-based Image or Video Coding," Data Compression Conference (DCC), pp.172-181, 2019.

[10] Perez-Daniel, R. Karina and V. Sanchez, "Luma-aware Multi-model Rate-control for HDR Content in HEVC," Proc. of IEEE International Conference on Image Processing (ICIP), pp. 1022-1026, 2017.

[11] G. J. Sullivan, J.-R. Ohm, W.-J. Han and T. Wiegand, "Overview of the High-Efficiency Video Coding (HEVC) Standard," IEEE Trans. on Circ. and Sys. for Video Tech., vol. 22, no. 12, pp. 1649-1668, 2012.

[12] E. Niedermeyer and F. Lopes da Silva, Electroencephalography: Basic Principles, Clinical Applications and Related Fields, Lippincott Williams & Wilkins, 2005.

[13] M. Dimaki, P. Vazquez, M. H. Olsen, L. Sasso, R. Rodriguez-Trujillo, I. Vedarethinam and W. E. Svendsen, "Fabrication and Characterization of 3D Micro and Nanoelectrodes for Neuron Recordings," Sensors (Switzerland), vol. 10, no. 11, pp. 10339–10355, 2010.

[14] S. J. M. Smith, "EEG in the Diagnosis, Classification and Management of Patients with Epilepsy," Journal of Neurology, Neurosurgery and Psychiatry, vol. 76, Suppl. 2, pp. ii2-ii7, 2005.

[15] S. Scholler, S. Bosse, M. S. Treder, B. Blankertz, G. Curio, K.-R. Müller and T. Wiegand, "Toward a Direct Measure of Video Quality Perception Using EEG," IEEE Trans. Image Process., vol. 21, no. 5, pp. 2619–2629, 2012.

[16] L. Lindemann and M. Magnor, "Assessing the Quality of Compressed Images Using EEG," Proc. of the 18th IEEE Int. Conf. Image Process., pp. 3109–3112, 2011.

[17] H.-C. Li, J. Seo, K. Kham and S. Lee, "Measurement of 3D Visual Fatigue Using Event-related Potential (ERP): 3D Oddball Paradigm," Proceedings of 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, pp. 213–216, 2008.

[18] L. Acqualagna, S. Bosse, A. K. Porbadnigk, G. Curio, K.-R. Müller, T. Wiegand and B. Blankertz, "EEG-based Classification of Video Quality Perception Using Steady State Visual Evoked Potentials (SSVEPs)," Jour. Neural Eng., vol. 12, no. 2, p. 26012, 2015.

[19] C. Tan, F. Sun and W. Zhang. "Deep Transfer Learning for EEG-based Brain Computer Interface," Proc of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[20] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J. Antons, K. Y. Chan, N. Ramzan and K. Brunnström, "Psychophysiology-based QoE Assessment: A Survey," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 1, pp. 6-21, 2017.

[21] S. Bosse, K. Müller, T. Wiegand and W. Samek, "Brain-Computer Interfacing for Multimedia Quality Assessment," Proc. of IEEE International Conference on Systems, Man and Cybernetics (SMC), Budapest, pp. 002834-002839, 2016.

[22] S. Bosse, L. Acqualagna, W. Samek, A. Porbadnigk, G. Curio, B. Blankertz, K. Mueller and T. E. Wiegand, "Assessing Perceived Image Quality Using Steady-State Visual Evoked Potentials and Spatio-Spectral Decomposition," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 8, pp. 1694-1706, Aug. 2018.

[23] F. S. Avarvand, S. Bosse, K. Mueller, R. Schaefer, G. Nolte, T. E. Wiegand, G. Curio and W. Samek, "Objective Quality Assessment of Stereoscopic Images with Vertical Disparity Using EEG," Journal of Neural Engineering, vol. 14, no. 4, p. 046009, 2017.

[24] L. Jia, Y. Tu, L. Wang, X. Zhong and Y. Wang, "Study of Image Quality Using Event-related Potentials Measurement," Jour. Electronic Imaging, vol. 27, p. 033046, 2018.

[25] L. Jia, L. Wang, Y. Tu and X. Zhong, "Studying the Effect of ROI on Image Quality Using ERPS," Proc. of the 2nd IEEE Advanced Information Management,Communicates,Electronic and Automation Control Conference (IMCEC), pp. 829-833, 2018.

[26] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, 2004.

[27] S. Bosse, K. Brunnström, S. Arndt, M. G. Martini and N. Ramzan, "A Common Framework for the Evaluation of Psychophysiological Visual Quality Assessment," Quality and User Experience, vol. 4, no. 3, [Online], Available: https://doi.org/10.1007/s41233-019-0025-5.

[28] S. Arndt, J. Antons, R. Schleicher, S. Möller and G. Curio, "Using Electroencephalography to Measure Perceived Video Quality," IEEE Jou. of Selected Topics in Signal Proc., vol. 8, no. 3, pp. 366-376, 2014.

[29] S. Scholler , S. Bosse , M. Treder , B. Blankertz , G. Curio , K. Muller and T. Wiegand, "Toward a Direct Measure of Video Quality Perception Using EEG," IEEE Transactions on Image Processing, vol. 21, no. 5, pp. 2619-2629, May 2012.

[30] S. Möller, S. Arndt, J. Antons, G. Curio, R. Schleicher and S. Scholler, "A Physiological Approach to Determine Video Quality," Proc. of IEEE International Symposium on Multimedia, Dana Point, California, USA, pp. 518-523, 2013.

[31] P. A. Abhang, B. W. Gawali and S. C. Mehrotra, Technological Basics of EEG Recording and Operation of Apparatus, Dec. 2016.

[32] Juri D. Kropotov, Quantitative EEG, Event-related Potentials and Neurotherapy, ISBN 978-0-12-374512-5 Academic Press, 2009.

[33] J. W. Britton, L. C. Frey, J. L. Hopp et al., Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children and Infants, American Epilepsy Society, ISBN-13: 978-0-9979756-0-4, 2016.

[34] R. B. Mfarij, Video Compression from EEG, M.Sc. Thesis, Department of Computer Engineering, Yarmouk University, Irbid, 2017.

ملخص البحث:

في هذه الدراسة، تستخدم إشارة تخطيط الدماغ من أجل تعديل معدل الضغط للفيديوهات بناءً على درجة اهتمام المشاهد. ويتم ذلك من خلال ربط معدّل الضغط الذي يخضع له الفيديو بدرجة اهتمام الشخص الذي يشاهده باستخدام مجموعة من الأُطُر. ويقاس اهتمام المشاهد بمجموعة من الأُطُر باستخدام إشارة تخطيط الدماغ من أجل بيان استجابة المشاهد للفيديو. وتُستخدم تقنيات إحصائية تطبق على إشارة تخطيط الدماغ لاستخراج الأُطُر موضوع الاهتمام بالنسبة للمشاهد. وقد تم تقييم الطريقة المستخدمة في هذه الدراسة قياساً على عدد من المتغيرات (أعلى نسبة إشارة الى ضجيج، ومؤشر التماثل البنيوي، والخطأ التربيعي المتوسط)، وذلك بمقارنة أداء التقنية المقترحة بتقنية (MPEG-4) التقليدية. وبينت النتائج خفضاً في حجم الفيديو بنسبة 15% مقارنة بتقنية (MPEG-4) دون الإضرار بجودة الفيديو.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 05, No. 03, December 2019.

A PROPOSED MODEL OF SELECTING FEATURES FOR CLASSIFYING ARABIC TEXT

Ahmed M. D. E. Hassanein[1] and Mohamed Nour[2]

(Received: 25-Jul.-2019, Revised: 15-Oct.-2019, Accepted: 2-Nov.-2019)

ABSTRACT

Classification of Arabic text plays an important role for several applications. Text classification aims at assigning predefined classes to text documents. Unstructured Arabic text can be easily processed by humans, while it is harder to be interpreted and understood by machines. So, before classifying Arabic text or documents, some pre-processing operations should be done.

This work presents a proposed model for selecting features from the adopted Arabic text; i.e., documents. In this work, the words 'text' and 'documents' are used interchangeably. The adopted documents are taken from Al-Khaleej-2004 corpus. The corpus contains thousands of documents which talk about news in different domains, such as economics, as well as international, local and sport news. Some preprocessing operations are carried out to extract the highly weighted terms that best describe the content of the documents. The proposed model contains many steps to define the most relevant features. After defining the initial number of features, based on the weighted words, the steps of the model begin. The first step is based on calculating the correlation between each feature and class one. Depending on a threshold value, the most highly correlated features are chosen. This reduces the number of chosen features. The number of features is again reduced by calculating the intra-correlation between the resultant features. This is done in the second step. The third step selects the best features from among those which resulted from the second step by adopting some logical operations. The logical operations, specifically logical AND or logical OR, are applied to fuse the values of features depending on their structure, nature and semantics. The obtained features are then reduced in number. The fourth step is based on adopting the idea of document clustering; i.e., the obtained features from step three are placed in one cluster. Then, iterative operations are used to group features into two clusters. Each cluster can be further partitioned into two clusters ...and so on. That partitioning is repeated till the clusters' contents are not changed. The contents of each cluster are fused together using the cosine rule. This reduces the overall number of features.

This work adopts four types of classifiers; namely, Naïve Bayes (NB), Decision Tree, CART and KNN. A comparative study is carried out among the behaviors of the adopted classifiers on the selected number of features. The comparative study considers some measurable criteria; namely, precision, recall, F-measure and accuracy. This work is implemented using WEKA and MatLab software packages. From the obtained results, the best performance is achieved by using CART classifier, while the worst one is obtained by using KNN classifier.

KEYWORDS

Text classification, Text clustering, Feature selection, Arabic datasets, Machine learning methods, Performance evaluation.

1. INTRODUCTION AND RELATED WORK

The majority of text classification research is directed to text written in English, while little research works have been carried out on Arabic text. There are hundreds of millions of people in twenty-two countries in Asia and Africa who speak Arabic as their native language. There are more than one billion Muslims who use Arabic during their prayer and reading the Holy Quran. So, more research is needed for classifying Arabic text to satisfy the requirements of Arabic text users. Text categorization or text classification plays an important role for a lot of applications. It is concerned with assigning labels to a set of documents, where such labels are known *a priori*. Examples of such applications include, but are not limited to: classification of news, email messages and web routing. Text classification can also be used in email routing, spam filtering, automated indexing of scientific articles, searching for information on the WWW, among others [1]-[2]. Many research efforts are exerted to classify Arabic text with high accuracy. Examples of such efforts include, but are not limited to the following research studies. Laila Khreisat [3] presented a research work an classifying

[1] A. M. D. E. Hassanein is with Systems and Information Department, Engineering Division, National Research Centre (NRC), Dokki, Giza, Egypt.. Email: ahmed.diaa.hassanein@gmail.com

[2] M. Nour is with Electronic Research Institute (ERI), Cairo, Egypt. Email: mnour99@hotmail.com

Arabic text documents. The author uses the N-gram frequency statistics employing dissimilarity measures; namely, Manhattan distance and Dice's measure of similarity [3]. A comparison is made to evaluate performance using the two adopted measures. The N-gram document classification using Dice’s measure outperforms that using the Manhattan measure [3]. Majed Ismail Hussien et al. [4] presented some text classification algorithms; namely, sequential minimal optimization (SMO), Naïve Bayes (NB) and J48. The algorithms are implemented using WEKA package and operated on Arabic text. A comparative study among the adopted algorithms is carried out focusing on classification accuracy, error rate and classification time as important measurable criteria [4]. A huge number of features lead to a bad performance in terms of both accuracy and time. During the implementation work, the SMO classifier achieved the best accuracy and lowest error rate, followed by J48, then the NB classifier [4]. The SMO algorithm proved to be the fastest one, followed by NB and then J48 classifier; i.e., the J48 classifier takes the highest amount of time [4]. Fadi Thabtah et al. [5] conducted the Naïve Bayesian algorithm based on chi-square feature selection method for categorizing Arabic data. The authors presented several experimental results compared against different Arabic text categorization datasets [5]. The study concluded that feature selection often increases classification accuracy by avoiding rare or non-significant features. Riyad Al-Shalabi et al. [6] evaluated the use of K-Nearest Neighbor (KNN) to classify Arabic text. The authors used a corpus which consists of more than six-hundreds of documents that belong to six categories. They implemented a method to extract keywords based on document frequency threshold (DF) methods [6]. The work achieved about 95% micro-average precision and recall scores [6]. KNN is good with small number of training patterns, provided that there is a sufficient number of examples for each category [6]. The selection of the feature space, the training dataset and the value of K can affect the classification accuracy. Jafar Ababneh et al. [7] stated that many text categorization approaches from data mining and machine learning exist. Examples of such approaches are: decision trees, support vector machine, neural networks, statistical methods, among others. The authors presented and compared the results obtained against Arabic text collections using KNN algorithm. Three different experiments are conducted on Arabic datasets. The experimental results operated on Saudi datasets revealed that cosine similarity outperforms both Disc and Jaccard coefficients. Anshul Goyal and Rajni Mehta [8] mentioned that classification is important with broad applications. It classifies each item in a set of data into one of predefined set of classes. The authors compared between the performance of Naïve Bayes (NB) and J48 classification algorithms [8]. NB is based on probability, while J48 is based on decision tree. The comparison took place using the context of a financial institute dataset to maximize true positive rate and minimize false positive rate rather than achieving higher accuracy. The authors used classification accuracy and cost analysis as measurable criteria [8]. The results showed that the efficiency and accuracy of J48 were better than those of the NB method [8]. Adel Hamdan Mohamed et al. [9] presented a method for Arabic text categorization using support vector machine, Naïve Bayes and neural networks. The authors mentioned that several research efforts were presented for classifying English text, while unfortunately few efforts were conducted on Arabic text classification. The authors analyzed and applied the classification methods mentioned above to classify Arabic data. A comparative study was carried out using a fixed number of documents for all categories of documents in training and testing. The results showed that the support vector machine is very promising [9]. Here, we aim to apply a different approach than those applied in previous works done, where the results of each step are analyzed and evaluated. According to the results of a previous step, a next step is proposed and applied.

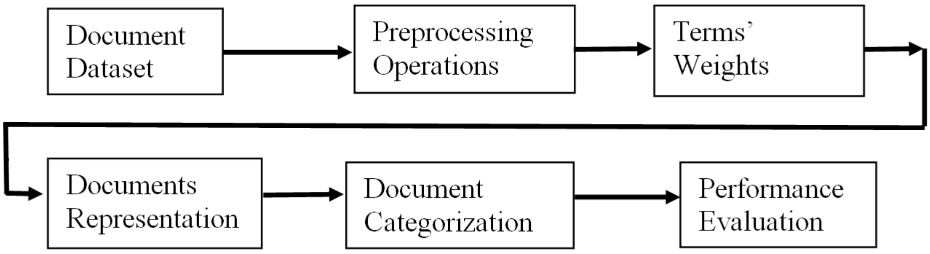


Figure 1. The main framework of text classification.

The main framework and building blocks for text or document classification are shown in Figure 1. The classification process involves several steps among which are: having a dataset and applying pre-processing operations, term weight computation, document representation and categorization and performance evaluation [10]-[15].

The dataset step is concerned with collecting the different documents that are to be classified. The pre-processing operations involve handling many themes, such as text tokenization, stop words' removal and word stemming. Document representation is important to reduce the complexity of documents and make them easy for handling. A document can be represented in a vector form using the vector space model; i.e., a document can be represented by a vector of words [12]. To perform document classification, documents are split into a training set and a testing set. The training set is used to build a model and make the system learn how to recognize different patterns of categories. The testing set is used to evaluate the system [6], [16]. Regarding the evaluation of the classification process, some measurable criteria are used. These criteria can be accuracy, precision, recall and F-measure [5]. In this paper, the same steps which are mentioned in Figure 1 are applied. The organization of this work is as follows: Section two presents an overview of the classification approaches which are used here. Section three conducts the Arabic dataset collection and the handling of some pre-processing operations. Section four introduces the proposed method for selecting the most important features for classifying the Arabic text. This step involves many themes of feature selection, such as correlation, intra-correlation, logical operations, clustering and fusion. Section five concludes the whole work and proposes possible future work.

2. Overview of Classification Approaches and Performance Evaluation

In this paper, four classifiers are adopted and investigated when classifying the dataset. The classifiers are the Naïve Bayes (abbreviated NB), Decision Tree, CART and KNN, respectively. Moreover, for the evaluation of the classification results, the measurable criteria used are accuracy, precision, recall and F-measure.

2.1 The Naïve Bayes Classifier

NB classifier works well with natural language processing (NLP) classifications. It is a supervised probabilistic algorithm that makes use of the probability theory and Bayes theorem to predict the class of a text.

NB classifier requires class conditional independence. This means that the effect of an attribute on a given class is independent of those of other attributes. If it is assumed that the training dataset $D=\{d_1,d_2......d_n\}$ contains n instances, each has a set of features and is represented as $d_i=\{x_{i1},x_{i2}......x_{in}\}$. The dataset D contains a set of classes $C=\{c_1,c_2,.....c_m\}$. Each training instance $d\in D$ has a particular class label c_i. The NB classifier predicts that an instance d belongs to a class c if and only if $P(c_i \mid d) > P(c_j \mid d)$ for $1\le j\le m, j\ne i$. The class c_i is the maximum posteriori hypothesis and it is the one for which $P(c_i \mid d)$ is maximized. The equation used is as follows [17]:

$$P(c_i \mid d)=\frac{P(d \mid c_i)\times P(c_i)}{P(d)} \tag{1}$$

where, $P(c_i \mid d)$ is the probability of document d to belong to class c. From the equation, $P(d)$ is constant for all classes, while $P(d \mid c_i)\times P(c_i)$ needs to be maximized. The class prior probabilities are calculated by $P(c_i)=\frac{|c_{i,D}|}{|n|}$, where $|c_{i,D}|$ is the number of training instances belonging to the class c_i in D and n is the number of the documents in the whole set.

2.2 The Decision Tree Classifier

The decision tree classifier is another type of supervised learning algorithms which is used in classification problems. For this algorithm, data is split into two or more homogeneous sets (or sub-populations) according to a certain splitter or differentiator in input variables. Splitting is done using several mathematical formulae. Entropy is one formula in which if p stands for success rate and q stands for failure rate, then reduction in Entropy is carried out by minimizing the formula [18]:

$$Entropy = -p\log_2(p) - q\log_2(q). \tag{2}$$

A second formula is the variance which is applied by reducing the equation [18]:

$$Variance = \frac{\sum\left(x-\bar{x}\right)^2}{n} \tag{3}$$

where, $\bar{x}$ stands for the mean of the values, x is the actual value and n is the number of values.
A third formula is Chi-square which is applied by maximizing the equation [18]:

$$Chisquare = \sqrt{\frac{(Actual - Expected)^2}{Expected}} \tag{4}$$

where, $Actual$ stands for the real class of a certain instance and $Expected$ stands for an expected class of a certain instance.

2.3 Classification and Regression Tree (CART) Classifier

CART stands for classification and regression tree which is one type of decision tree classifiers. It uses Gini method to create binary splits in a dataset. It is calculated for sub-nodes by using the sum of squares of probability for success and failure. If p stands for success rate and q stands for failure rate, the Gini formula is [19]: $\left(p^2 + q^2\right)$.

CART is important, as it deals with data using predicted and input features. CART can perform calculations and classification using both numerical and categorical parameters.

Gini index measures how well a given attribute classifies training samples into targeted classes. CART involves binary splitting of attributes, as it provides a hierarchy of univariate binary decision. The steps of CART are briefly mentioned as follows [19]: the first step is to know how the splitting attribute is selected. The second step involves setting the stopping rules and their application criteria. The third and last step is to decide on how nodes are assigned to classes.

2.4 The K-Nearest Neighbour (KNN) Classifier

The K-Nearest Neighbor belongs to the supervised learning algorithms. In this algorithm, we have a set of instances X, each having a group of features. Each instance belongs to one of a group of classes Y. The problem is to classify a new instance 'x' to one of the classes. The KNN algorithm can use several measures to define to which class or category a new instance belongs. Examples of such measures are Euclidean distance, cosine similarity, inner product similarity, among others. For the vectors of attributes (e.g. A and B), the Euclidean distance $d(A,B)$ is calculated using the following equation [20]-[22]:

$$d(A,B) = \sqrt{(a_1-b_1)^2 + (a_2-b_2)^2 + ...(a_n-b_n)^2}\ . \tag{5}$$

The k instances having the smallest distances to the new instances are grouped to choose the majority vote on the classes to which they belong. The class with the highest vote is the class of the new instance.

Some researchers use the cosine similarity in the KNN algorithm. The cosine similarity is a measure of similarity between the two vectors of n dimensions. That measure finds the cosine angle between the two vectors using the following formula [23]-[24]:

$$similarity = \cos(\theta) = \frac{A.B}{\|A\|\|B\|} \tag{6}$$

Other researchers use the inner product similarity which is known as the dot product or scalar product. It can be computed using the following formula [23]-[24]:

$$similarity = A.B \tag{7}$$

2.5 Measurable Criteria for Evaluating Performance

The performance of the proposed approach in terms of feature selection and the adopted classifiers are evaluated. To facilitate such evaluation, some measurable criteria are taken into consideration. Accuracy is one of the important themes in evaluating performance. Accuracy is defined as the ratio between the number of correctly identified documents and the total number of documents. Accuracy can be briefly expressed in terms of precision ($\Pr ec$), Recall ($\operatorname{Re} c$) and F-measure (FM), which are considered quantitative metrics. Precision ($\Pr ec$) can be defined as follows [25]-[27]:

$$\Pr ec = \frac{TP}{(TP + FP)} \tag{8}$$

Recall ($\operatorname{Re} c$) can be defined as follows [28]-[29]:

$$\operatorname{Re} c = \frac{TP}{(TP + FN)} \tag{9}$$

F-measure (FM) can be defined as follows [30]-[31]:

$$FM = \frac{2*(precision*recall)}{(precision + recall)} \tag{10}$$

where, TP is the number of documents which are correctly assigned to a certain category, FN is the number of documents which are not falsely assigned to a certain category, FP is the number of documents which are falsely assigned to a certain category and TN is the number of documents which are not correctly assigned to a certain category.

To determine the reliability of the proposed approach as well as that of the adopted classifiers, Arabic documents were taken as a test-bed. The researchers of this work selected a part of the documents in the corpus, not all of them. Very big-sized documents and very small-sized documents were avoided as explained in the next section.

3. Data Set Collection and Pre-processing Operations

We used Al-Khaleej-2004 corpus which contains more than 5000 Arabic documents. The corpus was taken from the website “https://sourceforge.net/projects/arabiccorps/”. The documents talk about daily news and are divided into four categories; namely, economy, international, local and Sport news. The average number of words and the average number of characters per word for each of the categories are calculated as shown in Table 1. The international category has the highest average number of words per document, but the lowest average number of characters per word. The economy category has the highest average number of characters per word. The lowest average number of words per document is the one calculated for the sport category. From the corpus, documents which have shooting numbers compared to the averages are rejected for our study. These documents were considered inaccurate representatives of the dataset to be selected. Including documents in the used dataset which have shooting numbers is avoided to overcome the problem of any error in the calculation of the average. The average is used in the calculation of the term weight, which is an important term in finding the other measurable criteria, such as precision and recall percentages. Equal numbers of documents from each category are chosen, so that the average numbers calculated for the whole corpora -as shown in Table 1- are still maintained for the selected ones.

Table 1. The average number of words per document and the average number of characters per word.

	Words/File	Characters/Word
Econ	461.92	6.17
Inter	561.89	5.96
Local	404.01	6.15
Sport	386.35	6.09

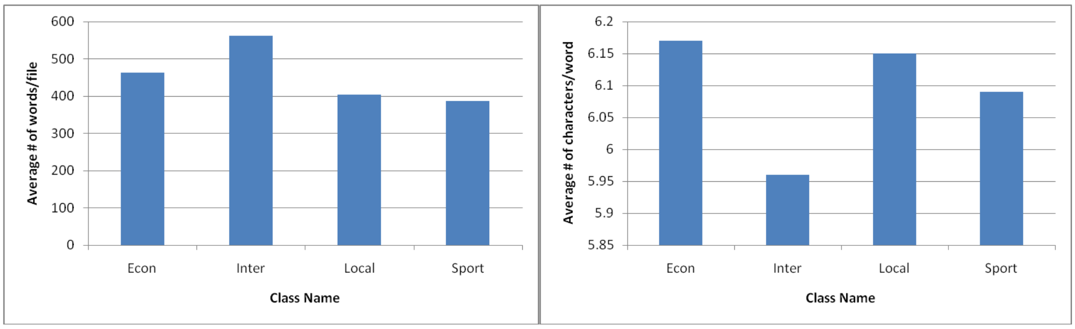


Figure 2. (a) # words/file in the dataset. Figure 2. (b) #characters/word in the dataset.

The results in Table 1 are plotted in Figure 2 in order to clarify the differences. In this research work, WEKA and Matlab software packages are used for the calculations. WEKA is a collection of machine learning codes that can be used for data mining tasks. It contains codes for data pre-processing, classification, regression, clustering, association rules and visualization. Matlab is a commercial software package with a specialized machine learning toolbox. The creation of our features is dependent on the term or word weight. Term weight computation can be performed using many methods among which are Information gain, Relative document frequency, Chi-square, Robertson 4th formula and Robertson 1st formula [11]. A term weight is assigned to each word or feature according to its frequency in each document. If the term frequency is high and appears in few documents, that term or feature is considered important to distinguish the document contents. In this paper, the term weighting is expressed as [11]:

$$w_{i,j} = tf\left(i, j\right) \times idf\left(i, j\right) = tf\left(i, j\right) \times \log\left(\frac{n}{df(j)}\right) \tag{11}$$

where, $w_{i,j}$ is the weight of the term j in document i, $tf\left(i, j\right)$ is the occurrence of term j in document i and $idf\left(i, j\right)$ is the inverse document frequency. $df\left(j\right)$ is the number of documents which contain feature j and n is the number of all documents in the dataset.

Next, we want to represent each document d_i, where i is the document number, in an array of a number of words w_n, where n is the number of words. The problem now is to investigate the minimum number of words or features per document which are sufficient to describe each document. The Naïve Bayes (NB) classifier is initially used to test the success of classification using a different number of words to describe each document. We start by choosing a single word with the highest frequency to represent each document, so $d_i = \left(w_1\right)$, and then increase the number of words per document. We aim to find out how the increase in the number of words per document will affect the classification accuracy.

Table 2. Precision, recall and f-measure for the four categories when using different numbers of words per document. Classification was carried out using the NB classifier.

		Prec	Rec	F-M			Prec	Rec	F-M			Prec	Rec	F-M
2 words/ document	Econ	71%	16%	26%	4 words/ document	Econ	78%	78%	78%	6 words/ document	Econ	79%	67%	72%
	Inter	57%	29%	38%		Inter	96%	84%	89%		Inter	77%	96%	86%
	Local	35%	79%	48%		Local	74%	78%	76%		Local	72%	61%	66%
	Sport	46%	35%	40%		Sport	92%	94%	93%		Sport	88%	97%	92%
	Avg.	53%	41%	38%		Avg.	84%	84%	84%		Avg.	79%	79%	79%

As shown in Table 2, when the number of words used to describe a document increases from two to four words per document, the average of the four categories for the three parameters; namely, precision, recall and f-measure increases by almost 100%. However, when the number of words increases from four to six words per document, the average for the three parameters decreases by almost 5%. So, the number of words which is selected to best describe each document is four words per document. The results are plotted in Figure 3 for all the calculations performed. Table 2 shows only three examples of the performed calculations. We can find that when we use 4 words per document, the accuracy of the classification is the best.

In Table 3, the confusion matrix for the four categories is shown. Positive true rates of classification are the highest for the four categories.

Table 3. The confusion matrix for the economy, international, local and sport news categories when using four words per document.

	Econ	Inter	Local	Sport
Econ	31%	1%	6%	1%
Inter	1%	22%	3%	0%
Local	6%	0%	28%	2%
Sport	1%	0%	1%	33%

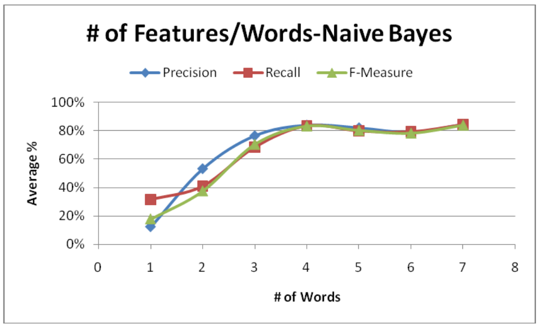


Figure 3. Measures for # words/document.

As we increase the number of words representing each document, the accuracy of the classification algorithm increases. As shown in Figure 3, we reach a point where increasing the number of words per document confuses the classification algorithm due to the repetition of words representing each document; i.e., the array of words representing each document is not any more unique for each document. Accordingly, the matrix which was fed to the WEKA software is created as follows:

1. For each document per category, the four terms with the highest term frequencies are selected to describe the content of the document.
2. For each category, we group the four words from each document for all the documents in the category and redundancy is removed.
3. For the four categories, all words that are repeated in the different categories were removed.
4. We have a matrix of 534 columns (words, attributes or features). The matrix fed to the WEKA is binary. If a word exists in a document, a one is placed in front of it; if not, a zero is placed.

4. Proposal of a Feature Selection Method

To our knowledge, 534 features are considered too large data to use in the classification problem. The time required to classify a document is large which will negatively affect the speed of calculations. Our proposal below involves many steps to reduce the number of features.

4.1 Correlation among Individual Features and Class Labels

The correlation between each of the features and the class column is calculated. Features which have the lowest correlations with the class are removed. Features are removed, so that the total number of features remaining to identify each document with its class decreases in increments of 25 features. Saying that two features have a high or low correlation with a certain class is a relative decision. A correlation value of 0.9 may be defined as high in one calculation and low in another calculation. This depends on other correlation values which we are comparing especially with the minimum and maximum values. That is why we don’t specify a threshold value to define features which are highly correlated or weakly correlated with a class. But, we chose to discard the 25 features with the lowest correlation values in every run as the threshold varies. We use the option of ten-fold calculations and the results are shown in Table 4. Figure 4 shows all calculations done, while Table 4 shows selected examples of the calculations.

In Table 4, the percentages of the precision, recall and f-measure decrease by almost 5% compared to those shown in Table 2. Precision, recall and f-measure percentages are stable as the number of attributes is decreased until reaching the knee point at 175 features. As one can see in Figure 4, we reduce the number of features from 500 to 25 and see the effect of this reduction on the accuracy of the results. When using less than 175 features in our classification, the percentages of recall and f-measure fall down. 175 features are the most efficient number to classify the documents to their corresponding classes with the highest possible accuracy. In our dataset, there exist features which are important in defining whether an instance belongs to a certain category or not. Removing features with the lowest correlations in groups of 25 features leads to the remaining of the features with have the highest correlations with the classes and which are detrimental in determining whether a document belongs to a certain class or not. Those remaining features are 175 ones. The confusion matrix when classifying documents using 175 attributes is shown in Table 5. It is shown that the highest error comes from classifying the documents under the local category. Next, we investigate why the local category is giving us the highest percentage of errors affecting all of our results later.

Table 4. Precision, recall and f-measure for the four categories when decreasing the number of attributes. The classification was done using the NB classifier.

		Prec	Rec	F-M
500 Attributes	Econ	78%	79%	79%
	Inter	94%	78%	85%
	Local	61%	74%	67%
	Sport	93%	86%	89%
	Avg.	81%	79%	80%

		Prec	Rec	F-M
175 Attributes	Econ	78%	79%	79%
	Inter	94%	78%	85%
	Local	61%	74%	67%
	Sport	93%	86%	89%
	Avg.	81%	79%	80%

		Prec	Rec	F-M
25 Attributes	Econ	93%	43%	59%
	Inter	96%	72%	82%
	Local	48%	91%	63%
	Sport	94%	82%	88%
	Avg.	83%	72%	73%

Table 5. The confusion matrix for the economy, international, local and sport news categories when using 175 attributes.

	Econ	Inter	Local	Sport
Econ	79%	1%	20%	0%
Inter	2%	78%	19%	1%
Local	16%	4%	74%	6%
Sport	3%	0%	11%	86%

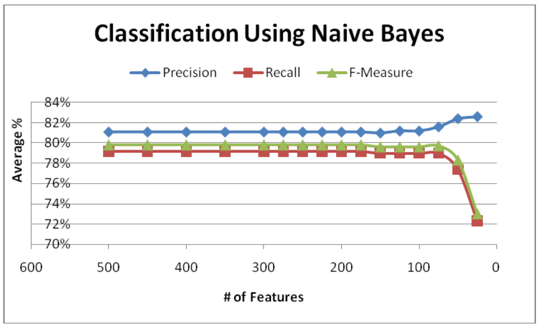


Figure 4. Measures using Naïve Bayes.

The aggregation of the four categories separately on a two-dimensional graph is examined. Two-dimensional reduction techniques are applied; namely, classical multidimensional scaling (CMDS) and a plot of the results for each two categories is shown in Figure 5. CMDS is a multivariate data analysis to explore the dissimilarity between the four classes which we have. From Figure 5, the aggregation of the points representing each category together against the three other categories is shown.

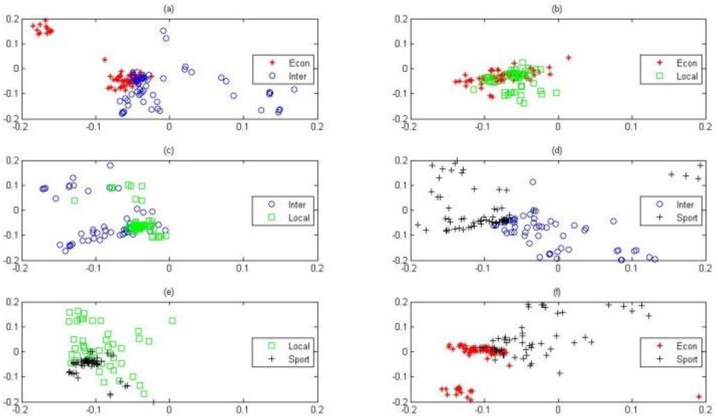


Figure 5. (a)-(f) A two-dimensional reduction in the dataset obtained for each two categories per graph. (a) Economy *versus* international, (b) Economy *vs.* local, (c) International *vs.* local, (d) International *vs.* sport, (e) Local *vs.* sport and (f) Economy *vs.* sport.

In Figure 5 (a), (d) and (f), the accumulation of the data points representing each of the economy, international and sport categories together is obvious. The separation of the data points of each of the three categories from the data points of the other two is clear as well. However, in Figure 5 (b), (c) and (e), the data points of the local category are randomly scattered through the data points of the other three categories, which makes the error in the identification of the documents belonging to the local category high. The differentiation and accordingly the classification of the documents under the local category from the other three categories have high percentage errors. This comes in accordance to what is shown in Table 4, which shows that the error is highest when classifying documents under the local category.

4.2 Operating the Adopted Classifiers on the Chosen Dataset

Different classifiers are adopted and operated on the dataset to see which kinds of classifiers can give better success rates of classification. Three classifiers; namely, decision tree, CART and KNN, are used, in addition to the Naïve Bayes (NB) classifier which has been used before.

Table 6. Precision, recall and f-measure using four different classifiers: NB, decision tree, CART and KNN classifiers.

		Prec	Rec	F-M
NB Classifier	Econ	78%	79%	79%
	Inter	94%	78%	85%
	Local	61%	74%	67%
	Sport	93%	86%	89%
	Avg.	81%	79%	80%

		Prec	Rec	F-M
Decision Tree	Econ	81%	45%	58%
	Inter	93%	59%	72%
	Local	42%	87%	57%
	Sport	93%	66%	77%
	Avg.	77%	65%	66%

		Prec	Rec	F-M
CART Classifier	Econ	81%	74%	77%
	Inter	89%	81%	85%
	Local	63%	81%	71%
	Sport	94%	84%	89%
	Avg.	82%	80%	81%

		Prec	Rec	F-M
KNN Classifier	Econ	76%	68%	72%
	Inter	84%	66%	74%
	Local	48%	77%	59%
	Sport	96%	64%	77%
	Avg.	76%	69%	70%

As shown in Table 6, the CART classifier shows the best results with 1% increase in precision, recall and f-measure compared to the NB classifier. For the CART and NB classifiers, the results of the three parameters for classifying documents under local category are first or second lowest, which is consistent with what was mentioned in Subsection 4.1. For the KNN and decision tree classifiers, precision and f-measure results for the local category are the lowest among all categories used. But, for the recall parameter, the Local category shows the highest percentage, which is inconsistent with all previous results. The calculation of the Recall parameter is inversely proportional with the false negative classification results. The recall parameter can be used in describing the success of our classification method, keeping in mind that false negative classification affects the results more than other measurable criteria. A comparison of the results is shown in Figure 6. The confusion matrix for the CART classifier is shown in Table 7.

Table 7. The confusion matrix using the CART classifier for the adopted categories.

	Econ	Inter	Local	Sport
Econ	74%	4%	21%	1%
Inter	3%	81%	15%	1%
Local	12%	4%	81%	3%
Sport	1%	2%	13%	84%

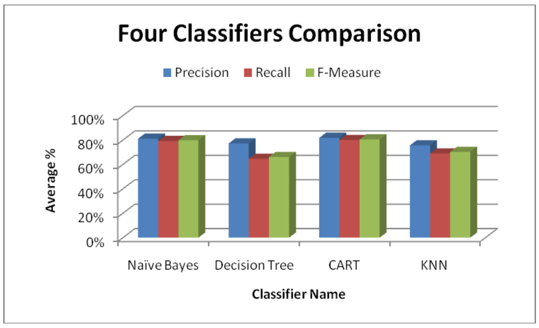


Figure 6. Average performance for the 4 classifiers used.

The classification of the documents under the local category shows better results than those for the NB classifier shown in Table 5. The error rate is lower and the success rate is higher. The CART classifier shows the best accuracy in the classification of our dataset among all classifiers used.

4.3 Intra-Correlation among Features

In this subsection, the total number of features is reduced for the whole dataset. The intra-correlation coefficients are calculated for all features. For each feature, features showing correlation values higher than a certain threshold are connected to it. Features with the highest number of connections are removed and the CART classifier is applied to see whether better classification results can be achieved or not. The threshold used here is chosen randomly to be 0.5. The number of connections is the measure to remove a feature or not. When a feature in question has a high number of connections, this means that there exist many other features which hold similar information to serve the accuracy of the classification method. We believe that removing this feature would not affect the accuracy of the classification results.

Table 8. Precision, recall and f-measure when reducing the number of features using correlation. Classification was done using the CART classifier.

		Prec	Rec	F-M
174 Attributes	Econ	81%	74%	77%
	Inter	88%	81%	84%
	Local	62%	78%	69%
	Sport	93%	84%	88%
	Avg.	81%	79%	80%

		Prec	Rec	F-M
142 Attributes	Econ	82%	73%	77%
	Inter	90%	78%	84%
	Local	59%	79%	68%
	Sport	96%	87%	91%
	Avg.	82%	79%	80%

		Prec	Rec	F-M
95 Attributes	Econ	78%	78%	78%
	Inter	94%	78%	85%
	Local	60%	74%	66%
	Sport	92%	85%	89%
	Avg.	80%	76%	77%

We aim to minimize the number of features to have percentages for the three parameters better than those shown in Table 7 or at least keep the percentages the same. As shown in Table 8, precision, recall and f-measure have the best results after removing features with the highest number of connections. As shown in Figure 7, the number of features is optimum before a drop down in the precision percentages is viewed (the knee point). The values for recall are stable until the knee point and after that, they fall down. Precision and f-measure values (81.7% and 79.9%, respectively) are the highest when the number of features is 142. Using a number of features higher than or lower than 142 decreases the accuracy of the classification method. Here, 142 is the minimum number of features which have intra-correlation values low enough so that none can substitute the existence of the others. The confusion matrix for the results of classification when using 142 features is shown in Table 9.

Table 9. The confusion matrix after removing the 17 connects using the CART classifier.

	Econ	Inter	Local	Sport
Econ	73%	2%	25%	0%
Inter	2%	78%	19%	1%
Local	13%	5%	79%	3%
Sport	0%	1%	12%	87%

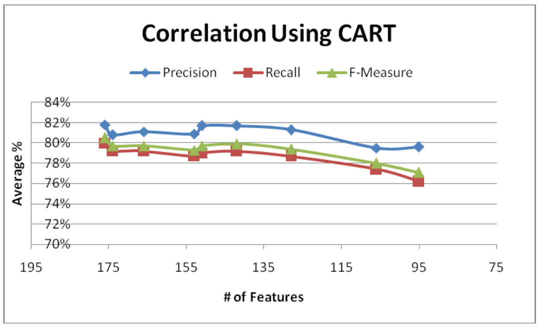


Figure 7. Measures using CART.

From Tables 7 and 9, we can see that after minimizing the number of features from 175 to 142, the accuracy achieved for the four categories is almost the same for the results of the classification percentages.

4.4 Bottom-Up Feature Fusion

Next, we apply the logical AND and OR operations to fuse the values of features together. The new features have no specific meaning, but they will reduce the total number of features. Each new feature will be the output of fusing the values of the two features into one.

Table 10. Precision, recall and f-measure when reducing the number of features using OR-binary operation. Classification is done using CART classifier.

		Prec	Rec	F-M
125 Attributes	Econ	84%	78%	81%
	Inter	94%	79%	86%
	Local	66%	81%	73%
	Sport	93%	92%	93%
	Avg.	84%	83%	83%

		Prec	Rec	F-M
75 Attributes	Econ	78%	74%	76%
	Inter	93%	87%	90%
	Local	66%	77%	71%
	Sport	92%	87%	89%
	Avg.	82%	81%	82%

		Prec	Rec	F-M
25 Attributes	Econ	69%	66%	67%
	Inter	88%	78%	83%
	Local	65%	71%	68%
	Sport	82%	86%	84%
	Avg.	76%	75%	76%

Inter-correlations are calculated between each two features. For the features with the highest correlation, we combine their values using logical OR. The number of features is decreased in increments of 25 features. The results of the classification are shown respectively in Table 10 and Figure 8. The best results are seen when reducing the number of features from 175 to become 125 features. The confusion matrix when using 125 features is shown in Table 11.

Table 11. Confusion matrix after reducing number of features to 125 using logical OR.

	Econ	Inter	Local	Sport
Econ	78%	1%	20%	1%
Inter	2%	79%	17%	2%
Local	12%	3%	81%	4%
Sport	0%	1%	7%	92%

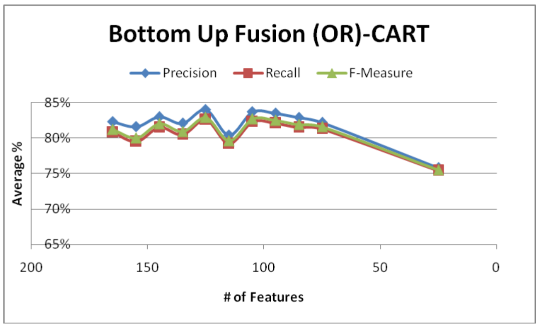


Figure 8. Measures using logical OR.

In Table 11, the results are better than those shown in Table 7. The percentage values of correct classifications for the local category are better. Next, the features are combined using the logical AND operation.

For the features with the highest correlation, we combine their values using logical AND. The number of features is decreased in increments of 25 features. The results of the classification are shown in Table 12 and Figure 9. The results are deteriorating as the number of features decreases. The confusion matrix when using 125 features is shown in Table 13.

Table 12. Precision, recall and f-measure when reducing the number of features using AND-binary operation. Classification is done using CART classifier.

		Prec	Rec	F-M
125 Attributes	Econ	59%	40%	48%
	Inter	72%	35%	47%
	Local	62%	34%	44%
	Sport	38%	86%	53%
	Avg.	57%	49%	48%

		Prec	Rec	F-M
75 Attributes	Econ	100%	5%	10%
	Inter	50%	2%	4%
	Local	83%	5%	9%
	Sport	27%	100%	42%
	Avg.	65%	29%	17%

		Prec	Rec	F-M
25 Attributes	Econ	24%	100%	39%
	Inter	0%	0%	0%
	Local	0%	0%	0%
	Sport	0%	0%	0%
	Avg.	6%	24%	10%

Table 13. Confusion matrix after reducing the number of features to 125 using logical AND.

	Econ	Inter	Local	Sport
Econ	40%	4%	11%	45%
Inter	4%	35%	6%	55%
Local	16%	7%	34%	43%
Sport	7%	2%	5%	86%

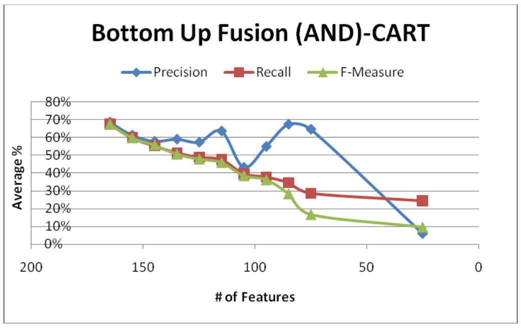


Figure 9. Measures using logical AND.

The results shown in Table 13 are the worst. The OR operation is the best method to combine different features together. Since our feature matrix is a binary one, when applying the OR logical operator on any 2 features, the output is a feature containing the information in both original features so that no information is lost. However; when applying the AND logical operator on the same 2 features, the output is a feature which only contains the shared areas so that the unshared data will be lost. That is why the classification results obtained when using the OR operator are much better than those obtained when using the AND operator.

4.5 Top-Down Feature Fusion

Next, features are combined together through unsupervised clustering. We start with all features in one cluster and then iterative methods are used to group features into two clusters. Each of the two clusters is further divided into two clusters …and so on. The results are shown in Table 14 and Figure 10.

K-means clustering is considered a partitioning algorithm. It can be used in several data mining tasks. It is considered a good algorithm to group a set of documents D into K groups or clusters. K-means clustering algorithm uses the maximum cosine similarity as a score for assigning a document to the more similar cluster centroid. The K-means algorithm is considered a proper algorithm to choose initial clusters' centroids. The document collection dataset D can be represented as $D=\left(d_1,d_2.......d_n\right)$, which can be grouped into k sets of coherent clusters. Moreover, each document d_i can be represented as a vector of weighted terms $d_i=\{w_{i1},w_{i2},......w_{it}\}$, where t is the number of all text features in D. For more details, the reader can refer to reference [11] and [21].

Table 14. Precision,recall and f-measure when reducing the number of features. Classification is done using the CART classifier.

		Prec	Rec	F-M			Prec	Rec	F-M			Prec	Rec	F-M
25 Attributes	Econ	74%	65%	69%	100 Attributes	Econ	83%	80%	81%	150 Attributes	Econ	82%	80%	81%
	Inter	80%	78%	79%		Inter	90%	81%	85%		Inter	92%	80%	85%
	Local	59%	71%	65%		Local	67%	80%	73%		Local	63%	76%	69%
	Sport	92%	86%	89%		Sport	94%	87%	90%		Sport	92%	86%	89%
	Avg.	76%	75%	75%		Avg.	83%	82%	82%		Avg.	82%	81%	81%

It is shown from Table 14 and Figure 10 that when features are distributed into 100 clusters, the classification gives the best results. The confusion matrix for classification using 100 clusters is shown in Table 15.

Table 15. The confusion matrix after reducing the number of features to 100.

	Econ	Inter	Local	Sport
Econ	80%	3%	17%	0%
Inter	2%	81%	15%	2%
Local	12%	4%	80%	4%
Sport	2%	2%	9%	87%

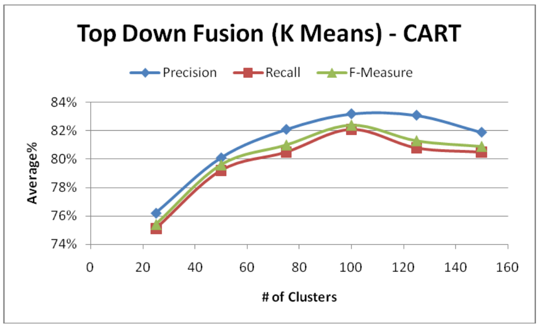


Figure 10. Measures for top-down feature fusion.

100 clusters are the best minimum number of features to classify the text under consideration. Distributing instances on clusters is grouping instances with shared values in one cluster so that at this point, fusing features under each cluster together is beneficial. But, when going further on increasing the number of clusters more than this point, the division of instances on clusters is not any more accurate and common grounds shared by instances are overstretched to the extent that two clusters may have almost similar features, which increases the classification error.

5. Comparing Proposed Method with Others

In this section, we compare the achieved results in this paper with the results of other feature reduction methods in two ways. The first is to use a standard feature reduction method found in WEKA and compare the accuracy of its classification results with those achieved in Table 15. The second is to compare our results with those of a previous work which used one of the state-of-the-art feature reduction techniques on the same dataset used here.

The "CfsSubsetEval" is one of the standard methods available in WEKA for feature reduction [26]. It is chosen randomly and applied to our dataset to compare the results with what we achieved in this paper. The chosen method evaluates the weight of each attribute based on its predictive ability for the class while minimizing redundancy in the final set of attributes selected. 38 attributes were selected by the "CfsSubsetEval" method [26] as the best subset of attributes that can give the highest accuracy in the classification problem. Then, the KNN classifier was applied as one of the standard available methods in WEKA for classification [26]. The results are shown in Figure 11.

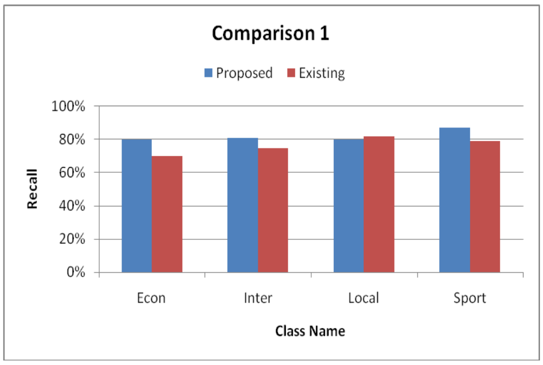


Figure 11. Comparing the achieved results with an existing standard work.

The proposed method achieved higher accuracy in the classification problem addressed, as shown in Figure 11. The proposed method produced 100 features which were needed to produce the results shown in the same figure. The accuracy of the proposed method is higher than that of the existing method for the three categories econ, inter and sport. For the local category, the proposed method still maintains a high accuracy, but slightly less than that of the existing method. As mentioned before in Section 4.1, the points representing the features of the local category are scattered through the other three categories and so, the accuracy of the identification of the documents under the local category has a higher percentage error.

In [32], one of the state-of-the-art methods of classification is used which is Multi-category Support Vector Machine (MSVM). The paper applies the method to the same dataset used here Al-Khaleej-2004. The accuracy of the results is almost equal to what we achieved with a difference of 1% more or less for the categories economy, international and sport news, as shown in Figure 12.

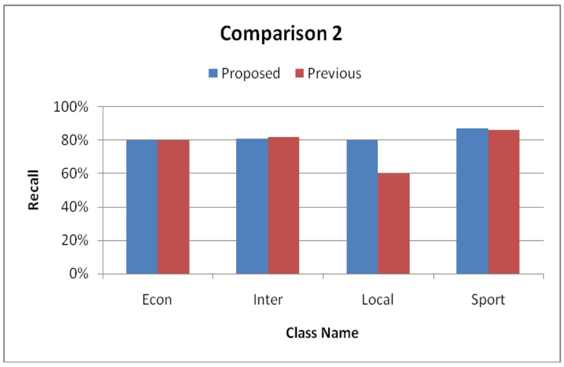


Figure 12. Comparing the achieved results with a previous work [32].

When comparing the recall values of the proposed work and the previous work for the four categories, it was noticed that the proposed work had the same, slightly better and much better values for the categories (econ and inter), sport and local, respectively, as shown in the figure. The properties of the features of the local category are discussed in Section 4.1 and we expected to have a larger error in the classification of its documents than in other categories. The proposed system succeeded in identifying the documents under the local category with almost a similar percentage error to that of the other three categories. The proposed system is resilient to problems which might exist in the used dataset than the MSVM method which was used in the previous work.

6. Concluding Remarks and Future Work

In this research work, the authors investigate and discuss the problem of Arabic text classification. Arabic documents are pre-processed by rejecting the stop words. Any document is tokenized into a set of words which are stemmed and weighted. The chosen weighted words are represented in a vector space or a feature vector. Four classifiers are operated on the chosen documents' feature vectors. The CART classifier is the best compared to the other adopted classifiers. The proposed feature selection approach improves accuracy, because it reduced the number of selected features. Precision, recall and f-measure are improved during the implementation of the steps of the proposed approach. The correlation between the individual features and the class labels, as well as the intra-correlation among the features played an important role in improving the classifier performance. Moreover, the fusion operations; either top-down or bottom-up, improve the performance of the classification process. This is clear from the values of precision, recall and f-measure, respectively. Such operations focus on selecting the most appropriate and significant features and ignoring the others. Finally, it is easy to say that the proposed feature approach can be applied on other datasets, because it is domain-independent.

Precision, recall and f-measure percentages are not as high as we would prefer to achieve. The study discussed here is based on word stemming in which the stem of all words is found and then term weighting is performed. Finding the stem of each word lacks a deeper view into the semantic relationship between the words used in each document. Including synonyms and antonyms of a word in the term weighting of the word would change the features of each document. Our proposed future work will focus on the semantics of Arabic words and how to include this deeper view into the selection of the features. New features may appear and already existing features may disappear, which can increase the accuracy of the classification method used.

References

[1] M. Suzuki, N. Yamagishi, T. Ishida, M. Goto and S. Hirasawa, "On a New Model for Automatic Text Categorization Based on Vector Space Model," Proc. of IEEE International Conference on Systems, Man and Cybernetics, pp. 3152-3159, 2010.

[2] R. Duwairi, "Arabic Text Categorization," International Arab Journal of Information Technology, vol. 4, no. 2, pp. 125-131, April 2007.

[3] L. Khreisat, "Arabic Text Classification Using N-Gram Frequency Statistics: A Comparative Study," Proc. of Conference on Data Mining (DMIN'06), pp. 78-82, 2017.

[4] M. I. Hussien, F. Olayah, M. Al-Dwan and A. Shamsan, "Arabic Text Classification Using SMO Naïve Bayesian, J48 Algorithms," International Journal of Recent Research and Applied Studies (IJRRAS), vol. 9, no. 2, pp. 306-316, November 2011.

[5] F. Thabtah, M. A. H. Eljimini, M. Zamzeer and W. M. Hadi, "Naïve Bayesian Based on Chi-Square to Categorize Arabic Data," Communication of the IBIMA, vol. 10, pp. 158-163, 2009.

[6] R. Al-Shalabi, G. Kanaan and M. Gharaibah, "Arabic Text Categorization Using KNN Algorithm,"[Online], Available: at the University of California Irvin data collections repository, http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.

[7] J. Ababneh, O. Almomani, W. Hadi, N. K. T. El-Omari and A. Al-Ibrahim, "Vector Space Models to Classify Arabic Text," International Journal of Computer Trends and Technology (IJCIT), vol. 7, no. 4, pp. 219-223, January 2014.

[8] Anshul Goyal and Rajni Mehta, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms", International Journal of Applied Engineering Research, vol. 7, no. 11, pp. 1-5, 2012.

[9] A. H. Mohamed, T. Alwada and O. Al-Momani, "Arabic Text Categorization Using Support Vector Machine, Naïve Bayes and Neural Networks," GSTF Jour. of Comput., vol. 5, no. 1, pp. 108-115, 2016.

[10] M. Labani, P. Moradi, F. Ahmadizar and M. Jalili, "A Novel Multivariate Filter Method for Feature Selection in Text Classification Problems," Eng. App. of Artificial Intell., vol. 70, pp. 25-37, 2018.

[11] L. M. Abualigah, A. T. Khader and E. S. Hanandeh, "A New Feature Selection Method to Improve the Document Clustering Using Particle Swarm Optimization Algorithm," Journal of Computer Science, vol. 25, pp. 456-466, 2018.

[12] Bhumika, S. S. Sehra and A. Nayyar, "A Review Paper on Algorithms Used for Text Classification", International Journal of Application or Innovation in Engineering and Management (IJAIEM), vol. 2, no. 3, pp. 90-99, March 2013.

[13] A. Elnahas, N. El-Fishawy, M. Nour, G. Attya and M. Tolba, "Query Expansion for Arabic Information Retrieval Model: Performance Analysis and Modification," Proc. of the Conference of Language Engineering, Cairo, December 6-7, 2017.

[14] S. A. Yousif, V. W. Samawi, I. Elkaban and R. Zantout, "Enhancement of Arabic Text Classification Using Semantic Relations of Arabic Wordnet," Journal of Computer Science, vol. 11, no. 3, pp. 498-509, 2015.

[15] M. M. Hijazi, A. M. Zaki and A. R. Ismail, "Arabic Text Classification: Review Study," Journal of Engineering and Applied Sciences, vol. 11, no. 3, pp. 528-536, 2016.

[16] S. Osama and M. Nour, "Feature Selection Methods for Predicting the Popularity of Online News: Comparative Study and a Proposed Method," Journal of Theoretical and Applied Information Technology, vol. 96, no. 19, pp. 6969-6980, October 15, 2018.

[17] D. Md. Farid, Li Zhang, C. M. Rahman, M. A. Hossain and R. Strachan, "Hybrid Decision Tree and Naïve Bayes Classification for Multi-Class Classifications Tasks," Journal of Expert Systems with Applications, vol. 41, pp. 1937-1946, 2014.

[18] A. Brunello, E. Marzano, A. Montanari and G. Sciavicco, "J48SS: A Novel Decision Tree Approach for the Handling of Sequential and Time Series Data," Computers Jour., vol. 8, no. 21, pp. 1-28, 2019.

[19] E. Venkatesan and T. Velmurugan, "Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification," Indian Journal of Science and Technology, vol. 8, no. 2, pp. 1-8, November 2015.

[20] Z. Elberrichi and K. Abidi, "Arabic Text Categorization: A Comparative Study of Different Representation Modes," International Arab Journal of Information Technology, vol. 9, no. 5, pp. 465-470, September 2012.

[21] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. Trippe and J. Gutierrez, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques, " KDD Bigdas, Halifax, Canada, pp. 1-13, July 2017.

[22] P. Kumbhar and M. Mali, "A Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification," Int. Jour. of Science and Research, vol. 5, no. 5, pp. 1267-1275, 2016.

[23] M. Abbas and K. Smaili, "Comparison of Topic Identification Methods for Arabic Language," Proc. of the International Conference of Recent Advances in Natural Language Processing (RANLP'05), Borovets, Bulgary, pp. 14-17, September 21-23, 2005.

[24] I. Rouby, M. Badawy, M. Nour and N. Hegazi, "Performance Evaluation of an Adopted Sentiment Analysis Model for Arabic Comments from the Facebook," Journal of Theoretical and Applied Information Technology, vol. 96, no. 21, pp. 7098-7112, November 15, 2018.

[25] N. Bhargava, G. Sharma, R. Bhargava and M. Mathuria, "Decision Tree Analysis on J48 Algorithm for Data Mining," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 6, pp. 1114-1119, June 2013.

[26] V. P. Bresfelean, "Analysis and Predictions on Students' Behavior Using Decision Trees in WEKA Environment," Proceedings of the 29th IEEE International Conference on Information Technology Interfaces, Croatia, June 25-28, 2007.

[27] T. R. Patil and S. S. Sherekar, "Performance Analysis of Naïve Bayes and J48 Classification Algorithms for Data Classification," International Journal of Computer Science and Applications, vol. 6, no. 2, pp. 256-261, April 2013.

[28] M. F. Zaiyadi and B. Baharudin, "A Proposed Hybrid Approach for Feature Selection in Text Document Categorization," International Journal of Computer and Information Engineering, vol. 4, no. 12, pp. 1799-1803, 2010.

[29] S. Francisca Rosario and K. Thangadurai, "RELIEF: Feature Selection Approach," International Journal of Innovative Research and Development, vol. 4, no. 11, pp. 218-224, October 2015.

[30] R. P. Durgabai, "Feature Selection Using Relief Algorithm," International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 10, pp. 8215-8218, October 2014.

[31] U. G. Mangai, S. Samanta, S. Das and P. R. Chowdhury, "A Survey of Decision Fusion and Feature Fusion Strategies for Pattern Classification," IETE Technical Review, vol. 27, no. 4, pp. 293-307, 2010.

[32] M. Abbas, K. Smaïli, and D. Berkani, "Multi-Category Support Vector Machines for Identifying Arabic Topics," Research in Computing Science, vol. 41, pp. 217-226, 2009.

ملخص البحث:

يلعب تصنيف النصوص باللغة العربية دوراً مهماً في العديد من التطبيقات. ويهدف تصنيف النصوص الى تخصيص أصناف معرّفة مسبقاً للوثائق النصية؛ إذ إن النصوص العربية غير المهيكلة ربما تكون سهلة المعالجة من جانب البشر، لكن فهمها وتفسيرها من جانب الآلة يكونان أكثر صعوبة. لذا، فإنه قبل تصنيف النصوص العربية، لا بد من القيام ببعض العمليات التي تدخل في باب المعالجة المسبقة.

يقدم هذا العمل البحثي أنموذجاً لانتقاء السِّمات من نصوص أو وثائق باللغة العربية. وهنا تستخدم كلمة (نصّ) وكلمة (وثيقة) بالمعنى ذاته. وقد تم استقاء النصوص لهذا العمل من (مجموعة الخليج)-2004. وتتضمن هذه المجموعة آلاف الوثائق التي تتناول أخباراً في مجالات مختلفة؛ مثل الأخبار الاقتصادية، والأخبار العالمية، والأخبار المحلية، وأخبار الرياضة. وقد تم إجراء عدد من عمليات المعالجة المسبقة من أجل استخلاص المفردات عالية الوزن التي تصف محتوى الوثيقة على النحو الأفضل. ويتضمن الأنموذج المقترح العديد من الخطوات من أجل تعريف السمات الأكثر ملاءمة. وبعد تحديد العدد الأولي من السمات، بناءً على الكلمات الموزونة، تبدأ خطوات الأنموذج المقترح. الخطوة الأولى مبنية على حساب الارتباط بين كل سمة من السمات والصنف الأول. وبناءً على قيمة عتبة محددة، يجري انتقاء السمات الأعلى ارتباطاً؛ الأمر الذي يقود الى تقليل عدد السمات المنتقاة. ثم يجري التقليل من عدد السمات مرّة أخرى عبر حساب الارتباط بين السمات الناتجة، ويكون ذلك في الخطوة الثانية. ومن خلال القيام ببعض العمليات المنطقية، يتم في الخطوة الثالثة انتقاء أفضل السمات من بين تلك التي نتجت من الخطوة الثانية، وذلك بناءً على عمليات (AND) و(OR) من أجل دمج بعض السمات تأسيساً على بِنْيتها وطبيعتها ودلالتها. وينجم عن ذلك تقليل آخر من عدد السمات. أما الخطوة الرابعة، فتقوم على فكرة (عَنْقَدة) النص أو الوثيقة، بحيث يتم وضع السمات التي نتجت من الخطوة الثالثة في مجموعة واحدة ومن ثم إجراء عمليات تكرارية تؤدي الى وضع السمات في مجموعتين، ليصار بعد ذلك الى تجزئة كل مجموعة من السمات الى مجموعتين ... وتستمر التجزئة الى أن تصبح محتويات المجموعات ثابتة لا تتغير. ويجري دمج محتويات مجموعات السمات باستخدام قاعدة جيب التمام؛ الأمر الذي يقلل العدد الإجمالي للسمات.

يستخدم هذا العمل أربعة مصنِّفات هي: (KNN, CART, DT, NB) للمقارنة بينها من حيث عدد السمات الناجم عن استخدام كل منها. وتأخذ الدراسة المقارنة بعين الاعتبار عدداً من المعايير، وهي: (F-M, R, P، إضافة الى دقة التصنيف). وقد تم تطبيق هذا الأنموذج باستخدام حُزم البرمجة: ويكا (WEKA)، وماتلاب (MATLAB). وتشير النتائج التي تم الحصول عليها الى أنّ المصنف (CART) كان الأفضل من حيث الأداء، بينما كان الأسوأ أداءً المصنف (KNN).

المجلة الأردنية للحاسوب وتكنولوجيا المعلومات (JJCIT) مجلة علمية عالمية متخصصة محكمة تنشر الأوراق البحثية الأصيلة عالية المستوى في جميع الجوانب والتقنيات المتعلقة بمجالات تكنولوجيا وهندسة الحاسوب والاتصالات وتكنولوجيا المعلومات.

تحتضن جامعة الأميرة سمية للتكنولوجيا (PSUT) المجلة الأردنية للحاسوب وتكنولوجيا المعلومات، وهي تصدر بدعم من صندوق دعم البحث العلمي في الأردن. وللباحثين الحق في قراءة كامل نصوص الأوراق البحثية المنشورة في المجلة وطباعتها وتوزيعها والبحث عنها وتنزيلها وتصويرها والوصول إليها. وتسمح المجلة بالنسخ من الأوراق المنشورة، لكن مع الإشارة إلى المصدر.

الأهداف والمجال

تهدف المجلة الأردنية للحاسوب وتكنولوجيا المعلومات (JJCIT) إلى نشر آخر التطورات في شكل أوراق بحثية أصيلة وبحوث مراجعة في جميع المجالات المتعلقة بالاتصالات وهندسة الحاسوب وتكنولوجيا المعلومات وجعلها متاحة للباحثين في شتى أرجاء العالم. وتركز المجلة على موضوعات تشمل على سبيل المثال لا الحصر: هندسة الحاسوب وشبكات الاتصالات وعلوم الحاسوب ونظم المعلومات وتكنولوجيا المعلومات وتطبيقاتها.

الفهرسة

المجلة الأردنية للحاسوب وتكنولوجيا المعلومات مفهرسة في كل من:

ELSEVIER Scopus

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

crossref Text and Data Mining

OCLC WorldCat

BASE

ACADEMIA

Google scholar

فريق دعم هيئة التحرير

المحرر اللغوي	ادخال البيانات وسكرتير هيئة التحرير
حيدر المومني	إياد الكوز

عنوان المجلة

الموقع الإلكتروني: www.jjcit.org
البريد الإلكتروني: jjcit@psut.edu.jo
العنوان: جامعة الاميرة سمية للتكنولوجيا، شارع خليل الساكت، الجبيهة، عمان، الأردن.
صندوق بريد: 1438 عمان 11941 الأردن
هاتف: +962-6-5359949
فاكس: +962-6-7295534

Princess Sumaya University for Technology
Since 1991 تأسست
PSUT
جامعة الأميرة سمية للتكنولوجيا

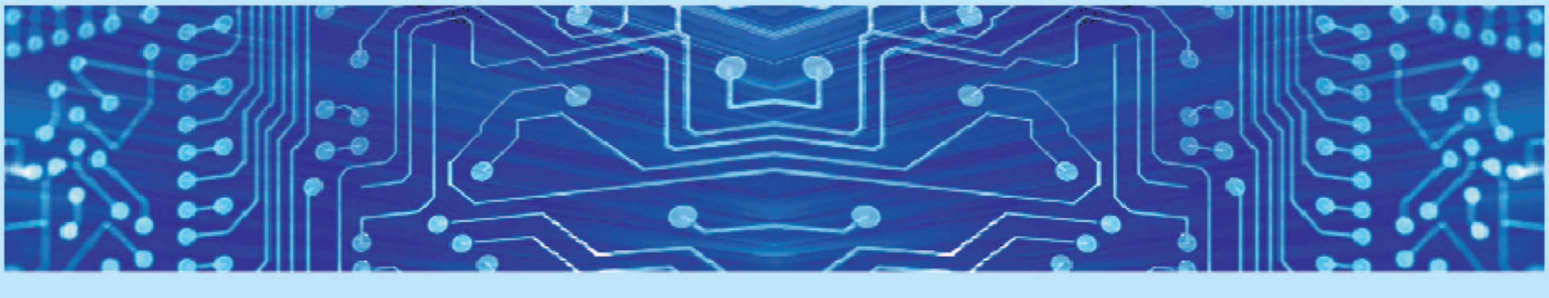
صندوق دعم البحث العلمي
Scientific Research Support Fund

المجلة الأردنية للحاسوب و تكنولوجيا المعلومات

كانون الأول ٢٠١٩ المجلد ٥ العدد ٣
ISSN 2415 - 1076 (Online)
ISSN 2413 - 9351 (Print)

JJCIT

jjcit@psut.edu.jo
www.jjcit.org



مجلة علمية عالمية متخصصة محكمة

تصدر بدعم من صندوق دعم البحث العلمي